DOCUMENT RESUME

ED 316 052 FL 018 375

AUTHOR Marty, Fernand; Hart, Robert S.
TITLE Computer Programs To Transcribe French Text into Speech: Problems and Suggested Solutions.
INSTITUTION Illinois Univ., Urbana. Language Learning Lab.
REPORT NO LLL-T-6-85
PUB DATE Mar 85
NOTE 223p.
AVAILABLE FROM Language Learning Laboratory, University of Illinois at Urbana-Champaign, G70 Foreign Languages Building, 707 South Mathews, Urbana, IL 61801.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC09 Plus Postage.
DESCRIPTORS Algorithms; Computational Linguistics; *Computer Software; *French; Microcomputers; *Phonetic Transcription; Programing; *Speech Synthesizers; Statistical Analysis; Structural Analysis (Linguistics)

ABSTRACT

The process of developing a computer program that can turn French text into speech is described. The program is designed to change a conventionally-spelled text into a phonemic transcription with an accuracy level of over 99.9%, divide the text into breath-groups that are syntactically correct and not over 14 syllables long, treat the unstable vowels and liaisons in accordance with the language level (register) chosen for the program, operate on personal microcomputers and use no more than 450 kilobytes of RAM, transcribe at a minimum rate of 10 phonemes per second, and, in combination with prosodic rules, be sent to a speech synthesizer. The strings of phonetic symbols generated can also be processed to yield statistical analyses, phonemic distribution tables, and graphs. The process of writing the computer program is described. Appended materials include: lists of words and sentences that can be used to test the program's accuracy; an algorithm for end-of-line hyphens; abbreviations written as complete words; treatment of hyphenated words; borrowings, neologisms, and slang; procedures for assignment of linguistic features; a list of linguistic features; root, prefix, and suffix tables; and algorithms, distributions, and procedures for various phonemes. A 53-item bibliography is included. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED316052

Language Learning Laboratory

LLL

LLL-T-6-85
MARCH 1985

The School of Humanities

University of Illinois
at Urbana-Champaign

COMPUTER PROGRAMS TO TRANSCRIBE FRENCH TEXT INTO SPEECH: PROBLEMS AND SUGGESTED SOLUTIONS

By Fernand Marty and Robert S. Hart

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Cheng, C. C.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

LANGUAGE LEARNING LABORATORY
School of Humanities
University of Illinois at Urbana-Champaign

Technical Report No. LLL-T-6-85

COMPUTER PROGRAMS TO TRANSCRIBE FRENCH TEXT INTO SPEECH:

PROBLEMS AND SUGGESTED SOLUTIONS

Linguistic Analysis by

FERNAND MARTY
Professor of French and Computer-based Education
University of Illinois, Urbana, Illinois 61801

Computer Programming by

ROBERT S. HART
Associate Director, Language Learning Laboratory
University of Illinois, Urbana, Illinois 61801

Copyright © 1985 by the Board of Trustees of the University of Illinois.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or other, without the prior written permission of the publisher.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the views or policies, expressed or implied, of the Language Learning Laboratory of the University of Illinois.

Available from: Language Learning Laboratory, University of Illinois at Urbana-Champaign, G70 Foreign Languages Building, 707 South Mathews, Urbana, Illinois 61801.

TO

DR. BRUCE H. MAINOUS

who, from 1972 to 1984,

directed the

LANGUAGE LEARNING LABORATORY

UNIVERSITY OF ILLINOIS

URBANA-CHAMPAIGN

TABLE OF CONTENTS

Acknowledgments vii

Foreword ix

I. PURPOSE 1

II. USES FOR OUR PROGRAM 1

III. PRESENT STATUS OF RESEARCH IN THIS FIELD 3

IV. A DIFFERENT APPROACH 4

V. PREPARATORY STEPS FOR WORD TRANSCRIPTION 5

STEP 1: Separation of text into strings with a maximum of 30 words 5
STEP 2: Checking that no word is over thirty characters long 9
STEP 3: Rewriting apostrophes 10
STEP 4: Rewriting mathematical characters 11
STEP 5: Miscellaneous rewritings 13
STEP 6: Hyphens at end of line 13
STEP 7: Treatment of abbreviations 13
STEP 8: Changing all remaining capital letters to lower-case letters 18

VI. TRANSCRIPTION OF INDIVIDUAL WORDS 19

A. WORD INPUT 19

B. THE LEXICON 19

1. Words 20
2. Idioms 20

C. THE ROOT TABLE 21

D. THE PREFIX TABLE 23

1. The need for a prefix table 23
2. Should the prefix table include all prefixes? 24
3. Unacceptable prefix separations 24
4. Combinations of prefixes and ambiguous spellings . . . 28
5. Prefixes other than those on our list 29

E. THE TABLE OF ENDINGS 29

F. REMOVING FINAL s 31

G. REWRITE FOR THE LETTER h 32

H. REWRITE FOR INTERVOCALIC s 32

I. SYLLABICATION 32

J. TRANSCRIPTION RULES 33

K. TRANSFORMATION RULES 47

VII. LINGUISTIC ANALYSIS OF THE INPUT STRING 52

A. RULES TO IMPROVE THE ASSIGNMENT OF LINGUISTIC FEATURES . . 52

B. RULES TO DISAMBIGUATE AMBIGUOUS SPELLINGS 59

C. FINAL STEP IN THE ASSIGNMENT OF LINGUISTIC FEATURES . . . 65

D. SEPARATION OF INPUT STRING INTO BREATH-GROUPS 65

E. PRONUNCIATION OF NUMBERS ENDING WITH 5, 6, 8, 10 69

F. ADDING THE CONSONANTS OF LIAISON 70

G. TREATMENT OF UNSTABLE VOWEL 76

H. RESULTS OF FINAL TEST 77

VIII. LEXICON 83

IX. APPENDICES 157

1. List of words which can be used to test the accuracy of a text-to-speech program 159
2. List of sentences which can be used to test the accuracy of a text-to-speech program 160
3. Algorithm for end-of-line hyphens 162
4. Abbreviations which are rewritten as complete words . . . 174
5. Treatment of hyphenated words (except at end of line) . . 179
6. Borrowings, neologisms, and slang 180
7. Procedures for the assignment of linguistic features . . . 182
8. List of linguistic features 187
9. Root table 190
10. Prefix table 191
11. Endings: modifications and assignment of features 193
12. Algorithm for initial /ˆ/ 194
13. Distribution of /a/ and /ɑ/ 196
14. The consonant /ɲ/ 199
15. Transcription of œ and æ 199
16. Algorithm for the unstable vowel 201
17. Description of linguistic features 77 through 112 204
18. Figures 208

X. FOOTNOTES 211

XI. BIBLIOGRAPHY 215

ACKNOWLEDGMENTS

We wish to express our appreciation to the Research Board of the University of Illinois (Urbana-Champaign) who supported this project for three years (1982-1985).

We also wish to thank the hundreds of French natives all over France who provided much valuable data for this research project.

FOREWORD

For the linguist and the computer programmer, designing a computer program that can read aloud with total accuracy printed materials such as newspapers, novels, and plays is a fascinating challenge.

On the phonemic level, attaining an accuracy level of about 95% is relatively easy; from that point on, while trying to reach 100% accuracy, one meets a multitude of problems, each affecting a very small linguistic domain and each requiring some complex programming. Human readers, because of their knowledge of the world, their grids of interrelated conceptual structures, and their ability to extract meaning, unsuspectingly solve problems of whose existence they are completely unaware and they can read on with graceful ease; indeed, it is doubtful that we humans could read if we had to go through the same process as a computer.

In a way, this is a report on irreconcilable differences, on what separates the human mind from its very own creation, the computer. This phonemic transcriber is only a modest contribution to the field of artificial intelligence, but it may be one of the frontier markers which separate what can be achieved from what cannot be achieved. Indeed, the nature of the problems we have encountered, the complexity of the solutions, and the time this project has required do suggest that natural languages, molded by chance as well as by necessity, do not readily fit the world of computers and that optimism as to the success of far more complex endeavors (e.g., optical scanners for handwritten texts, unrestricted speech recognition, corrective phonemic analysis of the pronunciation of foreign language students, corrective analysis of totally free written expression, fully automatic translation, or unrestricted dialogue with a computer) is probably not warranted.

COMPUTER PROGRAMS TO TRANSCRIBE FRENCH TEXT INTO SPEECH: PROBLEMS AND SUGGESTED SOLUTIONS

I. PURPOSE

The purpose of this research project is to write a computer program that will fulfill these three objectives:

1. Change a conventionally-spelled French text[1] into a phonemic transcription with an accuracy level greater than 99.9%, divide the text into breath-groups which are syntactically correct and not over fourteen syllables long[2], and treat the unstable vowels and liaisons in accordance with the language level (registre) chosen for this program[3]. The information about punctuation marks, initial capital letters, change of typeface (e.g., roman to italics, normal size to boldface), and underlined words is stored for later use with the prosodic program.

[Appendices I and II (pages 159-162) contain words and sentences which can be used to test the accuracy of a text-to-speech program.]

2. Operate on personal microcomputers and use no more than 450 kilobytes of RAM. [We expect that the prosodic rules will require about 40 additional kilobytes of RAM.]

3. Operate in real time, that is transcribe any French text at a minimum rate of ten phonemes per second (average speed used by a person reading aloud to another person).

As of March 1985 (publication date of this report), the components of this program have been tested on the mainframe computer of the Computer Education Research Laboratory at the University of Illinois. The phonemic accuracy level is over 99.9% and the transcription rate averages less than 0.005 second per phoneme.

We began to work on this program in 1974; we estimate that we need one more year to transfer the program to individual microcomputers. After the phonemic transcriber is operational on individual microcomputers, we will add the prosodic features (frequency, intensity, and length).[4]

II. USES FOR OUR PROGRAM

The strings of phonemic symbols released by our computer program can be processed to yield statistical analyses, phonemic distribution tables, and graphs. When combined with prosodic rules, they can be sent to a speech synthesizer, thereby producing a system that can pronounce any text typed at a keyboard, stored in a data base, or transmitted by an optical scanner.

Our program can, therefore, be used as follows:

1. Literary analysis:

With our program, it is possible to calculate the distribution of phonemes in a French text and print the corresponding graphs (see Figure 1, page 208). The input can come:

a. from an optical scanner: Some optical scanners (for example, the Kurzweil reader) can now identify printed characters in any typeface[5]. The output from an optical scanner can be sent directly to our program and the phonemic distribution can be calculated automatically.

b. from a computer tape or a data base: Many French literary texts, newspapers, magazines, etc. are now available in a computerized form which can be accessed by our program.

c. directly from a computer keyboard if an optical scanner or a computerized version of the text is not available; the text can be typed in its ordinary graphemic form.

2. Aid to the physically-handicapped:

a. Blind people have access to the texts which are available in Braille and to the texts which have been recorded on discs and tapes, but the amount of materials is insufficient. With text-to-speech programs, they can have access to the increasingly large amount of texts available in computerized form; with the addition of an optical scanner, they can have access to any printed material.

b. A text-to-speech program can be connected to a special typewriter so that a blind typist can hear each word and, if necessary, make corrections before the machine prints the word. It can also allow mute persons to voice their thoughts and, for example, exchange telephone conversations.

3. Teaching techniques in second-language acquisition:

We expect that the voice quality of speech synthesizers will become good enough to supplement the present tape and disc recordings in some aspects of second-language teaching. We will then be able to add the following improvements to computerized language courses:

a. In addition to receiving a written error analysis, the students could receive some <u>oral</u> feedback which would be especially useful in dictation and comprehension exercises.

b. Individualized audio exercises could be prepared according to the deficiencies of individual students.

c. Our program could constitute a pronouncing dictionary. The

students could type French sentences and hear them pronounced.

d. If the speech produced by voice synthesizers should become as good or nearly as good as human speech, we could save considerable time and money by storing the audio exercises in a phonemic/prosodic transcription and having them voiced by the synthesizer.[6]

III. PRESENT STATUS OF RESEARCH IN THIS FIELD

English and French are two of the languages which, because of the history of their spelling "systems", present enormous obstacles to the preparation of a completely automatic and unrestricted text-to-speech computer program; to our knowledge, the three objectives presented above in Part I have not been met yet. At present, we either find programs which attempt to reach high accuracy levels through the use of vast computer resources or programs which can run on individual microcomputers but have accuracy levels of only about 90% in the phonemic domain.

The first type of programs can have applications in some special areas with large financial resources (e.g., national defense). The second type is being used especially for the blind. One of the justifications for not providing higher accuracy (beside cost in time and money) is that the users are generally working with their native languages; in English, for example, they can understand the audio output without undue difficulty even though strings such as ea, ei, th are sometimes mispronounced and even though the program does not perform the linguistic and/or semantic analysis necessary for:

-the correct pronunciation of many words such as lives (verb or noun), wind (verb or noun), lead (verb or noun), tear (verb or noun), sow (verb or noun), house (verb or noun), read (present or past), longer (comparative or noun), bow (bow of a ship or bow used in archery), row (array or brawl).

-the correct stress placement in many cases such as present (verb or adjective/noun), object (verb or noun), duplicate (verb or adjective/noun).

-the shift of the primary stress in many cases such as pho'tograph, photog'rapher, photograph'ic, e'qual, equal'ity, equalitar'ian.

The following recent statements reflect the belief that a low level of accuracy is good enough:

> "Many blind persons use mediocre quality text to speech systems with great success. Nearly all blind persons who try find that with a few hours listening they can easily adjust for weird pronunciations -- it is like having a friend with a foreign accent."
> (Professor Peter Maggs, statement on the PLATO computer system of the University of Illinois, "micronotes", September 25, 1983)

> "Researchers at the Massachusetts Institute of Technology have taken the somewhat different tack of using the computer program to find linguistic structures.... The morphological table that forms the reference

dictionary of the program contains 12,000 feature entries... The program quickly becomes behemoth, requiring a large scale host computer with memory to spare. All this for voice output? It isn't worth it." (Teja E. R., 1981, p. 78)

And the authors of a French text-to-speech program write:

"Il ne pouvait être question de tenir compte des nombreuses exceptions aux règles de prononciation." (We did not attempt to take into account the numerous exceptions to the rules of pronunciation.)
(Maggs P., Trescases P., 1980, p. 232)

IV. A DIFFERENT APPROACH

What are the most effective ways of using computers in the field of linguistics? Should we attempt to give computers the kind of intelligence we have or should we find ways that best utilize the attributes of computers: their vast superiority in the manipulation of some types of information, but their marked inferiority in their ability to extract meaning?

In our view, it is a serious mistake —in the field of linguistics— to try to train a computer to duplicate the operation of the human mind. A computer cannot be taught to read in the same way as we, humans, were taught to read. The mental processes which we use to determine that, in a particular context, English present is to be stressed on the first syllable, that French président is a verb and not a noun, that English winds is a verb rather than a noun, that French fils means threads and not sons, etc. are too complex and are not sufficiently understood to be transferable to a computer.

Our efforts, during the past ten years, have been directed precisely at discovering the elements of information which computers can use effectively, information which is not stored in, or even perceived by, the human mind. For example, how many speakers of French know that, of all the words ending in emment, there is only one which is not an adverb and is not pronounced /amɑ̃/? All the information we present in our chapter about prefixes and ambiguous spellings is of such a nature; it is of such complexity that it is not directly useable by the human mind, but it is the kind of information which computers need and process with an efficiency and speed which cannot be matched by the human mind. Furthermore, computers process information in a manner which is quite different from the one used by the human mind (Figure 2, page 209).

The problems which human readers and computers encounter are often different; for example, a human reader will pronounce préparent without the slightest hesitation because he does not "see" that préparent can be interpreted as the prefix pré + the ambiguous spelling parent, but a computer will "see" that analysis and will need some special information in order to arrive at the correct transcription.

On the other hand, a computer —if properly programmed— will not have some of the problems which human readers encounter. There are words which

have a peculiar pronunciation and which we do not use often enough to always remember how to pronounce them; we have to consult a dictionary: archétype, catéchumène, cromlech, exarchat, faonne, fuchsine, nicotiane, syngnathe, unciforme, vergeure, etc. If we encounter les hamadryades, petits hayons, should we make the liaison or not? We may forget, but a computer will not forget; if the correct statements have been included in the program, the correct pronunciation will always be given.

The 99.9% accuracy level is obtained almost entirely through segmentation (prefixes, roots, endings) and transcription rules. The lexicon for words and idioms which cannot be treated correctly by segmentation and rules is relatively short (7399 items). One of the authors of this report (Fernand Marty) started preparing this lexicon in 1947 at Middlebury College, not —of course— for this text-to-speech project, but for his classes in French phonetics. The items in the lexicon have been collected from numerous treatises about French pronunciation and the thousands of books, newspapers, magazines, etc. which he has read. The lexicon entries include recent borrowings (e.g., bowling), slang (e.g., clebs), and also words which have been encountered only once or twice (e.g., catoblépas). The accuracy level would probably remain at 99.9% even if half those words were removed from the lexicon. A 100% level cannot be posited since, obviously, there are exception words which have not been met, since borrowings from other languages will continue (see Appendix VI, page 180), since some unusual abbreviations or acronyms might not be analyzed correctly (see page 13), since the lack of accents on capital letters might create some problems[7], since hyphens at the end of a line might sometimes be analyzed incorrectly (see Appendix III, page 162), and since —occasionally— a neologism might not be transcribed correctly by our computer program (see Apprendix VI, page 181).

V. PREPARATORY STEPS FOR THE TRANSCRIPTION OF INDIVIDUAL WORDS

STEP ONE: Separation of text into strings with a maximum of 30 words

A maximum must be set since specific amounts of computer memory must be allocated for each of the various tasks to be performed by the computer program. Thirty words is generous since the average number of words between true punctuation marks is about twelve. When, occasionally[8], we encounter more than thirty words without a punctuation mark, we use a sub-program which puts a string delimiter before some conjunctions and/or prepositions. This string of no more than 30 words, occurring between true punctuation marks, will be referred to as the input string. The punctuation marks indicating the end of the input string are retained since the information they provide is needed for the linguistic analysis and the prosodic features.

[Since the various rewritings which take place under STEP FOUR and STEP SEVEN might increase the number of words in the input string beyond our limit of 30 words, we have included code which, in such a case, would subdivide the input string; but we have not needed this additional code in any of texts we have tested.]

Many of the input strings are too long to constitute a breath-group. The division into breath-groups not over fourteen syllables long takes place after the grammar categories have been determined (page 65).

[Extremely rarely, one finds a text in which emphasis is expressed by placing spaces between the letters of a word (e.g., un vrai m i r a c l e); our program would treat these letters as separate words.]

Question marks, exclamation points, semicolons, dashes, parentheses, brackets, ellipsis marks[9], and one or more blank lines (e.g., after a title or headline) are always considered to be true punctuation marks. When several of these true punctuation marks occur together, only the last one is considered to be a punctuation mark. Example:

Input string 1: Tous les pays en faisaient partie [
Input string 2: sauf la Chine (
Input string 3: surtout cette année-là)];

Periods:

Periods are not treated as punctuation marks when they occur before a punctuation mark:

... vente de la nouvelle édit., des lettres ...

Periods are treated as punctuation marks:

1. when they occur at the end of a paragraph or text.

2. when they occur after another punctuation mark (the combination counts as only one punctuation mark):

... la fin du chapitre dix). Nous savons que ...

3. when they are followed by a space + a capital letter:

... la fin du chapitre dix. Nous savons que ...
... la fin du chapitre III. Nous savons que ...

except:

a. if the period is preceded by a single capital letter:

... la lettre pour A. Dubois ...
... la lettre de M. F. Dubois nous annonce que ...

except in: bombe A (H), jour J, heure H, degrés C (F), acte I (V), scène I (V, X), chapitre/livre I (V, X, L), système D, l'X, lettre/point + any capital letter, N. de D., and the specific cases listed in Appendix IV or in Group 5 of the lexicon (e.g., A preceded by a number, F preceded by a number):

... sans le système D. Oui, mais il faut avouer que ...

... qui sortait de l'X. Il déclara que ...
... la forme de la lettre M. Dubois remarque que ...
... du point A au point B. Cependant, il faut ...
... il crie N. de D. Pourtant ...
... qui coûte 27 F. Mais tout est plus cher ...
... de longitude O. Cependant le navire ...

b. or if the period is preceded by av, Av, boul, Boul, c.-à-d, cf, Ch, chap, ch.-l, Dr, éd, Ed, env, Env, ex, lat, MM, Ph, Pr, Prof, Pt, Sec, Th, viz:

... les choses vues boul. Marat par MM. Ch. Lidoy et Ed. Abert.
... en 300 av. J.-C. pendant plusieurs mois ...
... la capitale de la France, c.-à-d. Paris, a peu de parcs ...

Thus, periods are not treated as punctuation marks in: la S.N.C.F. dit, la R.a.t.p. est, art. 10, éd. de 1612, 34.812 francs, le lit d'A. Luc, deux années de prép. à Paris, le 28.5.1980, B.-du-R. ici.

[The above is not a complete solution, but the remaining problems are extremely rare. Problems may occur with strings of letters which can be words (e.g., cour, édit, long, sec) as well as abbreviations (e.g., le 12 cour., édit. de 1674, pour la long., une sec.) and for which it is not practical to set up conditions (as we do, for example, for art.) which can determine whether the string is a word or an abbreviation. We can be sure that strings like édit. are abbreviations only when the period has been determined not to be a true punctuation mark:

... votre lettre du 20 cour. nous est parvenue ...
... voici les prix de la nouvelle édit.: 25 francs ...
... dix mètres pour la long., cinq mètres pour la ...

When the period is followed by a space and a capital letter, we cannot perform the complex analysis that would determine whether such strings are words or abbreviations and we analyze the period as being a punctuation mark; this analysis is right for:

... mais il révoqua l'édit. Louis XIV déclara que ...

but it is wrong for:

... il préfère l'édit. Larousse à toutes celles que vous ...

Our program always analyzes Ed. as an abbreviation; in a sentence such as the following, the period would not be counted as a true punctuation mark and there would be no pause between Ed. and Je:

... qui s'appelle Ed. Je le connais depuis ...]

Commas:

Commas are treated as punctuation marks only before a space. Each comma

in the following examples is a true punctuation mark:

... cependant, Paul et Jeanne ...
... les numéros 321, 454, 789 ont gagné ...
... 13, rue des Pyrénées ...
... les lettres a, b, c, etc. sont ...
... qui a 50 m de long., 30 m de ...

but this comma is not a punctuation mark:

... la somme de 345,50 F payable ...

Colons:

Colons are treated as punctuation marks except when they occur between numeric characters (without spaces). The colons in the following examples are true punctuation marks:

... il a acheté: deux plats, trois verres, cinq ...
... dimensions pour les long.: 20 m pour le salon ...
... ont dépensé en 1980: 23 francs pour la peinture ...

but this colon is not a punctuation mark:

... 1234:23 donne un nombre ...

The procedure described above is necessarily complex since we must identify all the periods, commas, and colons which are not punctuation marks as in, for example:

La S.F.I.O. a dépensé 978,45 F à 12:15 h

Asterisks:

Asterisks occur only very rarely in the kind of texts for which this program is written. When they do occur, they usually indicate that a reference note is available, that a linguistic form is unattested, or that the author does not want to reveal the complete name of a person or place (e.g., Madame de B***). In our program, asterisks (single or multiple) are replaced with the word astérisque (see page 13).

Guillemets:

Guillemets (<< ... >>) constitute a special kind of punctuation mark; their use varies considerably in the kind of texts for which this program is written. There are books (mostly plays, novels, textbooks) without any guillemets at all; newspapers, on the contrary, are heavy users; in three randomly selected newspapers, we found an average of 30 sets of guillemets per page.

One of the uses of guillemets is to mark the beginning and end of a quotation; in addition, a different typeface may be used for the quotation within the guillemets (from roman type to italics, from italics to roman,

from regular size to boldface), but this usage varies considerably among the newspapers we have analyzed. Guillemets are also used for quotations within quotations. [However, Le Canard Enchaîné occasionally uses English quotation marks (e.g., <<En 1980, il disait "Je ne suis pas pressé d'être une vedette." Il continue de se hâter en prenant le temps de vivre.>>)]

> [We should note that guillemets are not the only way to indicate quotations; dashes at the beginning of a line can be used when a series of quotations is presented as a dialogue, or dashes can be used within a set of guillemets to indicate a change of speaker; occasionally, a change of typeface is used to indicate that the writer is using somebody else's words (e.g., Cela avait été de ma part un mouvement qu'il avait apprécié à sa juste valeur.)]

More than half the guillemets we found in our corpus were used for purposes other than indicating quotations. They were used with foreign words, with titles of books, with names of newspapers, ships, works of art, etc.; they were used to express irony, doubt, emphasis, a special meaning, or to indicate that a particular expression was used on purpose. In nearly half these cases, there was also a change of typeface. This use of guillemets also occurs within guillemets used for quotations (however, Le Canard Enchaîné then occasionally uses " ").

It is not always possible to determine the function of a particular set of guillemets (e.g., quotation or special use of a word). A critical reader cannot but be puzzled by the apparent absence of a logical system in the use of guillemets, italics, boldface, etc. On the very same page of a newspaper, the Soviet bloc is referred to as <<l'Est>> and l'Est; emphasis is expressed with guillemets without a change of typeface, with guillemets and a change of typeface, or with only a change of typeface; quotations are indicated with guillemets with and without a change of typeface.

Guillemets, therefore, are prosodic markers as well as phonemic markers; in our program, the words which are preceded by guillemets are assigned a linguistic feature (bit 76; see page 19). This feature is used to shape the prosody of the input string (e.g., pause, additional intensity), to prevent linking (for example, in il <<adorait>>, but not in l'<<amitié>>), to prevent liaison (e.g., quel grand <<idéologue>>), and to prevent the fall of the unstable vowel of monosyllabic words (e.g., pour ce <<beau>> décor). The unstable vowel of polysyllabic words is allowed to fall; then, linking is prevented (e.g., une telle <<idée>>) or a pause is inserted (e.g., une telle <<femme>>).

STEP TWO: Checking that no word is over thirty characters long

We allow a maximum of thirty characters per word (an accent counts as one character). We have not met in our readings any words longer than 30 characters (anticonstitutionnellement, which is said to be the longest French word, has only 25 characters). However, we have included in our code a sub-program which can handle words with over thirty characters; since any such word would begin with prefixes, we separate the prefixes and treat them as individual words: pseudo + anti + constitutionnellement.

STEP THREE: Rewriting apostrophes

The transcriber shown in Figure 2 (page 209) expects that the input string will not contain any apostrophes. However, removing all apostrophes (e.g., changing l'histoire to lhistoire) would not be satisfactory because:

1. We need the syntactic information provided by l', d', n', etc.

2. The lexicon entries beginning with a vowel or semivowel sound— instead of being listed only once— would have to be entered several times. For example, the verb eus would have to be entered also as jeus, meus, teus, leus, neus, queus.

Replacing the apostrophe with a space would create many ambiguities; for example, if le point d'attache became le point d attache, we would no longer know whether d stands for the isolated letter pronounced /de/ or the preposition pronounced /d/.

We have adopted for each occurrence of an apostrophe the solution which seems the most appropriate. For example, l'as, d'as, m'as, n'as, t'as, qu'as are changed to liy as, diy as, miy as, niy as, tiy as, quiy as; the combinations with iy are unambiguous, they can receive the necessary feature markers, and —at the proper time— can be transcribed as /l/, /d/, /m/, /n/, /t/, /k/, etc.

The following changes apply to capital letters and lower-case letters:

replace l' with liy + space	
replace j' with jiy + space	
replace m' with miy + space	
replace n' with niy + space	
replace t' with tiy + space	
replace s'il with sil	for s'il and s'ils
replace s' with siy + space	
replace c' with ciy + space	
replace ç' with ciy + space	
replace grand' with gran + space	for old spellings grand'mère, etc.
replace prud' with prud	for prud'homme and derivatives
replace rd'hui with rdui	for aujourd'hui
replace d' with diy + space	
replace presqu' with presk	for presqu'île
replace squ' with sk + space	for puisqu', jusqu', lorsqu'
replace quoiqu' with quoik + space	
replace lqu' with lk	for quelqu'un(e)
replace qu' with kiy + space	for all other qu'
replace mp' with mpp	for spellings champ', etc.
remove remaining apostrophes	entr'aider, v'là, v'lan

Throwing away the apostrophe in s'il(s), prud'homme and derivatives, aujourd'hui, presqu'île, quelqu'un(e) does not cause a loss of syntactic information. Spellings such as manip' (for manipulation), manif' (for manifestation), sup' (for supérieur) are transcribed correctly without the apostrophe by our rules; a few words like mam'zelle, sciences nat' would

have to be in the lexicon whether the apostrophe is dropped or not.

[Sometimes, the visual form is clearer (e.g., une partie de pok') or provides more information (e.g., chers z'électeurs) than the spoken form.]

STEP FOUR: Rewriting mathematical characters

As already stated, our program is designed to handle newspaper articles, novels, plays, textbooks dealing with subjects not requiring specialized symbols or characters (e.g., history, geography, literature); therefore, Greek letters (e.g., π) and symbols such as square root cause the program to stop and a message is sent to the user. The following manipulations are more than enough to keep our accuracy level above the 99.9% objective.

1. The characters "space + $" and "space + £", when directly in front of numeric characters, are moved to after the numeric characters and a space is added in front of the numeric characters. For example:

... la somme de $75 et ... becomes: ... la somme de 75 $ et ...

2. Superscript numbers after capital letters are changed to their graphemic form and spaces are added: SO^4H^2 becomes S O quatre H deux.

3. Superscript 2 and superscript 3, when they occur after numbers, are rewritten; for example, 15^2 is rewritten 15 puissance deux and 225^3 is rewritten 225 puissance 3.

4. The "words" 1^{er}, 1^{re}, I^{er}, I^{re}, V^e, X^e, 2^o, 3^o, 4^o, 8^o (whether the letters are superscript or not) become premier, première, premier, première, cinquième, dixième, secundo, tertio, quarto, octavo (except when º is followed by C or F). The remaining º following a number without a space are changed to degrés.

5. Roman numerals, composed of at least two characters (e.g., II, XIX), are changed to the Arabic system. For example, IX, III^e become 9, 3^e.

6. The single letters V and X are rewritten cinq and dix when they follow a word beginning with a capital letter. For example, Charles X becomes Charles dix.

7. I is rewritten un when it follows acte, scène, chapitre, chap., or livre; V is rewritten cinq when it follows acte, scène, chapitre, chap., or livre; X rewritten dix when it follows scène, chapitre, chap., or livre; L is rewritten cinquante when it follows chapitre, chap., or livre.

8. Numbers followed by e (whether superscript or not) are spelled out (without final e or s, and with final f changed to v) and ième is added: 6^e is rewritten sixième, 12^e is rewritten douzième, and 200^e is rewritten deux centième. [Our transcription rules transcribe cinqième correctly.]

9. The character - is replaced with a comma and space when it occurs between numbers if the input string also contains one of these strings:

phon, appe, num, compo. Example: 16-23-89-64 beomes 16, 23, 89, 64.

10. In a string "number, comma, space, number, character -, number", the character - is replaced with "space à space" if the last number is greater than the preceding number. Example: 19, 34-41 becomes 19, 34 à 41.

11. The character x is rewritten multiplié par when it occurs between numbers (with or without spaces).

12. The character : is rewritten divisé par except that the colon is replaced with a comma if the second number is followed by h. or h + space or h + end of input string. Example: 12:15 h becomes 12,15 h.

13. The character = is rewritten égal.

14. The characters + and - are rewritten pluss and moins when they occur before a number (with or without a space). [The spelling pluss is used to bypass the linguistic analysis which is required for the spelling plus.]

15. When a fraction sign occurs (for example, in $\frac{5}{6}$), our program reads the top number as a cardinal number; the lower number is read as demi if it is 2, tiers if it is 3, quart if it is 4, and as an ordinal number if it is greater than 4 (applying the same algorithm as in paragraph 8 above).

16. A string containing number smaller than 32, slash, number smaller than 13, slash, number is assumed to represent a date. The first and third numbers are read as cardinal numbers; the second number is replaced with the corresponding name of the month.

Other slashes occurring between numbers are replaced with sur: 5/8 is read as cinq sur huit. [This treatment of slashes is not entirely satisfactory since some newspapers, occasionally, use number/number to refer to a date (e.g., 16/11 could mean seize novembre).]

17. The strings mm^2, cm^2, dm^2, m^2, hm^2, km^2, cm^3, dm^3, and m^3 (whether 2 and 3 are superscript or not) are rewritten as millimètre carré, centimètre carré, décimètre carré, mètre carré, hectomètre carré, kilomètre carré, centimètre cube, décimètre cube, and mètre cube.

18. Other occurrences of superscript numbers are assumed to be references to footnotes and are replaced with "voir note" + graphemic form of superscript number.

19. The strings 7bre, 8bre, 9bre, Xbre are rewritten septembre, octobre, novembre, décembre.

20. If the sequence "number comma number" is followed by one of the abbreviations which on pages 175-179 are rewritten when they are preceded by a word having bit 74 set to 1 (ch, cl, F, etc.), we replace the comma with the abbreviation. For example, "124,35 cal." becomes "124 cal. 35".

21. All remaining numbers are changed to their graphemic forms and we set feature "rwnb" (bit 74) of each word to 1. For example, 2 954 280

becomes deux millions neuf cent cinquante-quatre mille deux cent quatre-vingts, 224,315 becomes deux cent vingt-quatre virgule trois cent quinze, 007 becomes zéro zéro sept, 12.45 becomes douze point quarante cinq[10], un beau 14-juillet becomes un beau quatorze-juillet.

A space is added when the number is not preceded by a space:

... nous écoutions TF1 sur l'autoroute A23 ...

becomes: ... nous écoutions TF un sur l'autoroute A vingt-trois ...

All remaining occurrences of the character - are hyphens between alphabetic characters (e.g., prenez-en) and are treated later.

STEP FIVE: Miscellaneous rewritings

The following string replacements simplify the transcription rules and the linguistic analysis (for example, the word beaux-fils is rewritten bofiss before the string is processed by the transcriber). The changes are as follows:

replace	[ghe], [ghi] with [gue],[gui]	for ghetto, malpighi, etc.
replace	[avant-hier] with [avantièr]	
replace	[petits-fils] with [petifiss]	
replace	[beaux-fils] with [bofiss]	
replace	[lock-out] with [lockout]	
replace	[é-je] with [è-je]	for aimé-je, etc.
replace	[s/] with [szx]	for s/sol, s/Dives, etc.
replace	[%] with [poursan]	
replace	[$] with [dolar]	
replace	[£] with [livre sterling]	
replace	[&] with [é]	
replace	[œ] with [oe]	for cœur, œil, etc.
replace	[æ] with [é]	for cæcum, pæan, etc.
replace	[*] (single or multiple) with [astérisque]	

STEP SIX: Hyphens at end of line

Hyphens occurring at the end of a line present considerable problems. End-of-line hyphens are discussed in Appendix III, page 162.

STEP SEVEN: Treatment of abbreviations

The term abbreviation designates shortened forms of words (e.g., art. for article, Mlle for Mademoiselle), forms with superscript letters (e.g., Dr for Docteur), and sequences of the initial letters of words (e.g., USMC for United States Marine Corps). The term acronym is used when the resulting combination of initial letters is pronounced as a word (e.g., radar from radio detecting and ranging, smig from salaire minimum interprofessionnel garanti); in many cases (e.g., Gestapo, igame, laser,

Simca, SAMU), most people no longer know or never knew the origin of the acronym. Nowadays the process is often reversed; first, the firm selects a few catchy words (historical, mythological, etc.) and then determines which one has letters which in some way fit the description of their product (e.g., ANTIOPE for Acquisition Numérique et la Télévisualisation d'Images Organisées en Pages d'Écriture).

A given abbreviation may represent different combinations of words. For example, depending on the context, F.M. may stand for franchise militaire, fusil-mitrailleur, force multinationale, or franc-maçon.

Depending on the author and/or publisher, a given abbreviation may appear with or without periods, with capitals only, with a capital only for the initial letter, or with only small letters. The presence or absence of periods does not indicate how the abbreviation should be pronounced: C.A.P.E.S. and SMIG are acronyms, but S.F.I.O. and RATP are pronounced letter by letter. In newspaper headlines written entirely in capital letters, periods can be used to prevent ambiguities with abbreviations such as AS, BIT, CAP, CAPET, CES, CET, CHU, CODER, DOT, OS, SI, TU. Thus, a newspaper headline will appear as GREVE DES O.S. A SOCHAUX rather than as GREVE DES OS A SOCHAUX, but the absence of periods does not matter if there is no difference of pronunciation (e.g., the acronyms FANE, SAC and the words FANE, SAC).

The average person is familiar with only a few hundred of the thousands and thousands of abbreviations which are in use today[11]; however, upon meeting an unknown abbreviation, the human reader can often give the right pronunciation. The combination of letters in some abbreviations (e.g., M.R.A.P.) is such that only a letter-by-letter pronunciation is possible; there are combinations of letters which can be pronounced as a word, but which do not sound ear-pleasing, or which may be confusing, or which may not match the image of the entity they stand for (e.g., C.H.U., D.U.P.) and for which, therefore, a letter-by-letter pronunciation is preferable; there are some abbreviations (e.g., A.F.N.O.R.) which are clearly meant to be acronyms; and there are some abbreviations for which both pronunciations are used. Under such conditions, it is extremely difficult to write a computer program which could match our linguistic "sense".

An analysis of over one thousand abbreviations[12] collected from a large number of newspapers and magazines published between 1944 and 1984 led us to write the following rules:

1. Abbreviations which are pronounced as if they were complete words:

Appendix IV (pages 174-179) contains the list of the abbreviations which —at this point in our computer program— are changed to their complete graphemic form or, in order to go faster through the rules, to a simplified form; except for N, S, E, O which are in the group of ambiguous spellings, we have also placed in Appendix IV the single letters which, under some conditions, are not pronounced as alphabet letters (e.g., g becomes gramme when preceded by a number). Examples:

art. when followed by a number is rewritten article

dép. when followed by a number or by À is rewritten départ; in other cases, it is rewritten département
Hme is rewritten homme
JF is rewritten jeunefille
etc. is rewritten ètsétéra
c.-à-d. is rewritten sétadire
Cte is rewritten comte
T.T.C. is rewritten toutetakskomprize
F preceded by a number is rewritten fran
B.-du-R. is rewritten boushdurône
M. after a space or if initial of input string is rewritten monsieur
[The strings lettre M. and point M. do not constitute a problem since, in such cases (see page 6), the period is a punctuation mark.]

Thus, the strings:

... À son collègue de l'O.N.M. M. Jean Divay ...
... 10 000 F T.T.C. payés comptant ....

become:

... À son collègue de l'O.N.M. monsieur Jean Divay ...
... 10 000 fran toutetakskomprize payés comptant ....

The problems unsolved by Appendix IV occur so rarely that they have no statistical significance for our program. However, in order to better understand the present limitations of computer programs in the handling of natural languages, it is important to list some of those problems:

a. abbreviations representing words other than those listed in Appendix IV: We rewrite div. as divers since this is the only meaning we have found in our corpus, but —in some special texts— div. could stand for dividende, division, or divorce. At the present time, we would have no way of deciding which rewriting should be used.

b. abbreviations not listed in Appendix IV: Some examples are:

-- P.O. P.Q. P.S. moy.

Those are examples of abbreviations which we have found in our corpus with several meanings. For example, most French people would pronounce petites ondes when P.O. refers to a radio station in the AM band; they would pronounce Pyrénées-Orientales when P.O. refers to a city in that département. Our program does not attempt the enormously complex context analysis which would be necessary to resolve this kind of ambiguity; P.O. is treated as in section 4 (page 18). The abbreviation P.Q. can stand for premier quartier, province de Québec, or parti québécois; P.S. can stand for parti socialiste or post-scriptum; moy. can stand for moyen or moyenne.

-- cour. édit. long. sec.

The problems posed by such strings were discussed on page 7.

-- Paris SG / RC Paris / JDD / PPDA / MD

People who are interested in soccer know that Paris SG and RC Paris are to be pronounced Paris Saint-Germain and Racing-Club de Paris. Most French people today know that JDD is the Journal du dimanche; they also know that PPDA and MD stand for the TV and radio personalities Patrick Poivre d'Arvor and Michel Drucker. Our program reads all such abbreviations as single letters; there is little or no value in making up a list since not all French people read such abbreviations as complete words and since most of these abbreviations remain in use for only a short time.

-- s/

s/ occurs as an abbreviation of sous (e.g., s/sol) and sur (e.g., Saint-Pierre s/Dives). This particular abbreviation is on the idiom list.

-- F. Mitterrand / A. Briand / R. Poincaré / Ph. Pétain

In newspapers, the use of last names with abbreviated first names is rare. In the first two pages of Le Canard Enchaîné of August 8, 1984, we found 158 last names: in 97 cases, only the last name was mentioned; in 60 cases, the last name was accompanied by the full first name; the first name was abbreviated in only one case (R. Treno). In the first two pages of Le Monde of August 9, 1984, we found 71 last names: in 23 cases, only the last name was mentioned; in 47 cases, the last name was accompanied by the full first name; the first name was abbreviated in only one case (J.-M. Colombani). On the other hand, some encyclopedias and dictionaries often abbreviate the first names; for example, on page 280 of QUID (1981), 77 out of 121 last names are given with the abbreviated first names.

Most French people, upon meeting an abbreviated first name, read it as a full first name if they know what the abbreviation stands for; Ch. de Gaulle, today, is certainly read as /ʃarldəgol/ rather than as /seaʃdəgol/. If they do not know what the abbreviation stands for, they either skip it or pronounce it as an alphabet letter.

We could have added to our lexicon a list of famous people with their abbreviated first names, but we decided not to do it since this list would have had to include thousands of names and since the lack of such a list does not interfere in any way with the operation of our program.

Except for the specific cases listed in Appendix IV, our program pronounces A., B., C., etc. as single letters; Ed. is in Appendix IV; Ch., Ph., and Th. are on the idiom list.

-- N. de D.

Abbreviations of this type have been placed on the idiom list.

-- Chilly-Min

Some abbreviations are understood only in small geographical

areas; for example, it is only in the southern suburbs of Paris that people know that Chilly-M^in stands for the city named Chilly-Mazarin. In all such cases, our program ignores the superscript and reads the abbreviated word as if it were a real word (/ʃijimɛ̃/).

-- all., prép., Méd., etc.

Practically, any word can be abbreviated. In some texts (e.g., technical reports), high-frequency words are abbreviated after they have been used once or twice: allumettes becomes all., préparatifs becomes prép., Médecine becomes Méd., etc. In other texts, all. could stand for alliage, prép. could stand for préposition, and Méd. could stand for Méditerranée. A search to determine which word is being abbreviated would not necessarily succeed; for example, in the following text, we would have great difficulty in writing a program which would establish that all. stands for allumettes and not for allemande, allocation, or even the verb allouer; our program would have to "know" that, of the four words which begin with all, allumettes is the one most likely to come in boxes:

... C'est la Kommandantur allemande qui décide comment allouer les stocks d'allumettes... une allocation de 300 boîtes contenant chacune 80 all. sera distribuée. Dans les bureaux ...

And, of course, our program could not even detect that an abbreviation is being used in:

... Il y aura une distribution de 300 boîtes contenant chacune 80 all. Dans les bureaux ...

In all such cases (fortunately extremely rare), our program treats the abbreviations as if they were whole words: /al/, /prep/, /me/.

2. Sequences of consonant letters occurring without intervening periods

We assume that such sequences are abbreviations; at this point in our computer program, we insert a blank space in between the consonants and, if capital letters are used, we replace the capital letters with lower-case letters. Examples:

ccp à Paris ...	is changed to: c c p à Paris ...
un HLM de banlieue ...	is changed to: un h l m de banlieue ...
un Hlm de banlieue ...	is changed to: un h l m de banlieue ...
un hlm de banlieue ...	is changed to: un h l m de banlieue ...
le pH de ce milieu ...	is changed to: le p h de ce milieu

3. Sequences of capital letters without intervening periods occurring in texts not written entirely in capital letters:

We assume that such sequences are abbreviations[13]; at this point in our computer program, we add the string ww in front and we change the capital letters into lower-case letters:

Réunion des AC à Ax ... is changed to: Réunion des wwac à Ax ...

du plan ORSEC qui ... is changed to: du plan wworsec qui ...
le SMIG passe à ... is changed to: le wwsmig passe à ...

These modified strings are handled by our transcriber as explained in section 5 below.

4. Sequences of letters separated by intervening periods:

We assume that such sequences are abbreviations; at this point in our computer program, we add the string ww in front, we remove all periods (except periods which mark the end of an input string), and we change the capital letters into lower-case letters:

la S.F.I.O. à Paris ... is changed to: la wwsfio à Paris ...
la R.a.t.p. en grève ... is changed to: la wwratp en grève ...
le c.r.o.u.s. de Paris ... is changed to: le wwcrous de Paris ...
le S.M.I.G. passe à ... is changed to: le wwsmig passe à ...
le S.m.i.g. passe à ... is changed to: le wwsmig passe à ...
le s.m.i.g. passe à ... is changed to: le wwsmig passe à ...

These modified strings are handled by our transcriber as explained in section 5 below.

5. Our program assumes that words prefixed with ww are to be pronounced letter by letter. Those which are not (e.g., wwsmig, wwafat, wwdeug, wwecu) and those in which e is pronounced /e/ instead of /ə/ (e.g., wwaef) have been added to our lexicon (group 6 and group 7) and are transcribed during the first pass through the lexicon. Immediately after this first pass, the remaining feature markers (ww) are removed and a space is inserted between the letters:

la wwsfio à Paris ... is changed to: la s f i o à Paris ...

Those individual letters, like those for the abbreviations in section 2 above, are transcribed correctly by our transcription rules.

6. Abbreviations which are used without intervening periods and which contain small letters are treated like ordinary words (e.g., Midem, Orsec, ovni, Smig, smig); those which cannot be transcribed correctly by our transcription rules (e.g., Crous, Afat) and those which are pronounced letter by letter have been placed in group 8 of the lexicon.

STEP EIGHT: Changing all remaining capital letters to lower-case letters

All remaining capital letters are changed to lower-case letters and bit 75 (feature "cap") of each word beginning with a capital letter is set to 1. We also set bit 18 (feature "prnn") to 1 if the capital letter occurs after a comma, a colon, a semicolon, a dash, a parenthesis, guillemets, or when it occurs inside an input string. This information is needed to assign additional linguistic features, to solve the problems posed by ambiguous spellings, etc.

An input string such as:

Le maire de Paris annonce l'adhésion de madame Lise Duc au r p r

becomes:

le(75) maire de Paris(18)(75) annonce l'adhésion de madame Lise(18)(75) Duc(18)(75) au r p r

[This algorithm does not catch all proper names; in particular, it fails to mark proper names occurring at the beginning of a paragraph or after a true period. For a further discussion, see page 187.]

[For the problems posed by unaccented capital letters, see footnote 7.]

VI. TRANSCRIPTION OF INDIVIDUAL WORDS

Each word in the input string is transcribed with IPA symbols. For this task, a dictionary look-up would not be practical; there are thousands of French roots in current use and we would have to include all the forms of each verb (41 forms for être, 36 for the so-called regular verbs of the chanter type) and all nouns, adjectives, etc. would have to be entered with all their masculine, feminine, singular, and plural forms.

The scheme we have adopted is shown in the flowchart of Figure 2 (page 209); Figure 3 (page 210) traces the transcription of a word. We will now describe in detail each of the steps in the chart.

A. WORD INPUT

As described in Part V (Preparatory steps for the transcription of individual words), some words and abbreviations may have been rewritten and the linguistic features "cap" (bit 75) and "prnn" (bit 18) may already have been assigned to some words as explained in Step 8 (page 18).

At this point in our program, we assign the features "sxzend" (bit 59; words ending with s, x, or z), "dtend" (bit 60; words ending with d or t), "nend" (bit 61; words ending with n), "rend" (bit 62; words ending with r), "gend" (bit 63: words ending with g), "quote" (bit 76; words preceded by guillemets), "prehyph" (bit 72; words preceding a hyphen: joues-en), "posthyph" (bit 73; words following a hyphen: joues-en), and all hyphens are replaced with a space (e.g., quatre-vingt-dix-huit is changed to quatre vingt dix huit). [This includes occurrences such as "for-mi-da-ble", "La petite-femme-qui-aime-bien-les-bêtes" (Grevisse (1980), page 153).] The replacement of hyphens with spaces is discussed in Appendix V, page 179.

B. THE LEXICON

The lexicon (pages 83-156) contains 7399 individual words and idioms. [Neologisms, borrowings, and slang items are discussed in Appendix VI.]

Words or idioms found in the lexicon are transcribed and the linguistic features are assigned as described on pages 184-186; when the spelling is ambiguous, the feature "ambi" (bit 71) is set to 1 to indicate that the word must go through the disambiguation process (page 59).

1. Words

There are eight groups of words:

a. The first group (1819 items) is composed of words which are exceptions to the classifications set up in our list of 95 endings (page 193). For example, the list contains:

- gemment because it is the only French word ending with emment which is not an adverb.

- acquiescent, immiscent because they are the only French words ending with scent which are not nouns.

- brut, rut because they are the only French words ending with rut which are not verbs.

- affront, amont, dont, front, mont, pont, rodomont because they are the only French words ending with ont which are not verbs.

b. The second group (1223 items) is composed of words (mostly high-frequency words) which facilitate the linguistic analysis of the input string: femme, homme, jeune, vieux, nouvel, ont, longtemps, etc.

c. The third group (2032 items) is composed of words which would not be transcribed correctly by our rules. This list includes ennemi, chaos, atlas, nerf, béotien, paonneau, oyez, football, dolmen, sirop, sud, toast, chrétienté, femmelette, etc.

d. The fourth group (816 items) contains proper names which our rules would not transcribe correctly (e.g., Desmoulins).

e. The fifth group (158 items) contains spellings which have several pronunciations (e.g., inventions) and which require some analysis.

f. The sixth group (186 items) contains the abbreviations which have the "ww" feature marker and which are not pronounced letter by letter.

g. The seventh group (8 items) contains the abbreviations which have the "ww" feature marker and in which the letter E is pronounced /e/.

h. The eighth group (29 items) contains the abbreviations which occur without periods and with small letters and which are not transcribed correctly by our transcription rules.

2. Idioms

The idiom list (287 main entries and 841 sub-entries) contains groups

of words with a special pronunciation:

- idioms with a type of liaison not normally made (e.g., dos à dos).

- consonant with a different pronunciation (e.g., neuf ans).

- consonant not pronounced (e.g., chef-d'oeuvre).

- idioms in which ambiguities are resolved (e.g., compte-fils, m'as-tu-vu, fier-à-bras).

It also contains groups of words which operate as function words: jusqu'à ce que, si bien que, etc.

When a word or idiom is found, it is transcribed with IPA symbols, the proper procedure for the assignment of linguistic features is followed (see Appendices VII and VIII), and the program goes on to the next word.

As shown on Figure 2 (page 209), a word may go through the lexicon table several times. Immediately after the first pass, the markers (ww), which indicate an abbreviation, are removed and a space is inserted between the letters:

wwsfio is changed to s f i o

C. THE ROOT TABLE

The next step, if a word is not found in the lexicon, is to determine whether the root should be rewritten. We have established a table of 228 roots; for example, the strings déségr, présupp, and rect are changed to déçégr, préçup, and rèkt. The root list is in Appendix IX, page 190.

The main reason for having a root table is that there are cases when we must prevent prefix separation. As shown in Figure 2 (page 209), a prefix search is not conducted after a root search has succeeded. Therefore, by placing parasit and désign on the root table, we ensure that the prefixes para and dés will not be separated in those cases and that parasite will come out with /z/ (intervocalic s) and that désigniez will not have /gn/ as it would if prefix separation allowed igni to become initial.

The most troublesome initial string (40 entries out of 228 in the root table) is re: the letter e can represent an unstable vowel which can be dropped (e.g., ressembler, reflet, regarder), an unstable vowel which is usually retained (e.g., rechigner, recrue), a closed /e/ (e.g., ressayer), an open /ɛ/ (e.g., rectangle), or —in combination with m or n— the nasal vowel /ɑ̃/ (e.g., rembourser, renivrer).

Note that the spelling of the root does not have to be changed; it is sufficient that the string be found on the root table to prevent prefix separation.

The root table is also used in conjunction with the transcription rules to solve some particular problems; for example, accourr and acquerr are

changed to acourr and aquerr so that rr will be transcribed as /rr/ rather than /r/ as in pourrez or verrez.

Another reason (see Figure 2, page 209) is that, once a root has been rewritten, the transcription process is faster since no further search through the lexicon is made. We must, however, keep the following in mind when preparing the lexicon:

a. Words which are exceptions to the categories set up in the root table must be found in the lexicon during the first pass. This means that the feminine and plural forms must also be listed when they exist: archal, assentiment(s), auxquels, auxquelles, auxerrois(e)(s), bienheureux, bienheureuse(s), biennal(e)(s), biennaux, chorizo(s), eurêka, renne(s), ressuions, ressuiez, trachée(s), trichons, trichions, trichiez, sprinter(s), etc.

b. If the transcription problem is in the part of the word which follows the root, we must also place all the forms of the word in the lexicon: aiguillier(s), archéosocial(e)(s), archéosociaux, archéosocialisme, ignipuncture(s), respect(s), etc.

c. Placing a root on the root table does not necessarily solve the problems of the root itself; for example, the presence of flirt, saoul, shoot, sprint, view on the root table provides accurate transcriptions for flirter, flirteur, saouler, saoulerie, shooter, sprinter, intervieweur, all the forms of the verb interviewer, etc., but the words flirt(s), saoul(s), shoot(s), sprint(s), interview(s) must be placed in the lexicon since the final t of flirt(s), shoot(s), sprint(s) is pronounced, since the l of saoul(s) is not pronounced, and since interview(s) ends with a vowel sound. Note that, since we do not go through the lexicon after a root has been rewritten, we must include the singular and plural forms of all such words in the lexicon.

d. We must be careful not to use a root rewriting which would result in a wrong linguistic analysis. For example, if patien were rewritten paçian, patient would be marked adjective/noun/present participle and patiente would be marked adjective.

We have placed some roots (e.g., bapt, compt, fais) in the transcription rules instead of placing them in the root table because those roots might occur with prefixes which are not on our list of prefixes (e.g., surcomptabilité). Our program is very flexible; it provides several avenues for the transcription of each word and, in each case, we have attempted to choose the avenue which yields the correct transcription at the lowest cost in time and computer processing power.

The roots must appear in the right order on the table. For example, restat (as in restatuer) must precede resta (as in restant) since, with the reverse order, restat would be rewritten incorrectly.

When rewriting the roots, we chose the combinations of characters that go through the transcription rules as quickly as possible. For example, we replaced s with ç whenever possible since the transcription

rules handle ç more rapidly than s.

We also had to rewrite the strings in ways that would yield the correct transcriptions; for all the roots in which we wanted the letter e to be transcribed /ɛ/, it was simple to change e to è (e.g., rept is changed to rèpt) since è is always transcribed /ɛ/. But what about the few roots which require that initial en be transcribed /ɑ̃/ (e.g., enivre) and the few roots which require that gn be transcribed /gn/ (e.g., diagnostic)? We use sequences of letters which do not occur in French: wxz for gn and wzx for en and we place the corresponding rules in unit w (page 47).

Our distinction between prefix and root is somewhat artificial; there are cases when what is obviously a prefix must be placed on the root table in order to solve a particular problem. For example, placing archa and archéo on the root table allowed us to solve the problems presented by words like archange.

Since each root (in our PLATO program) is limited to 10 characters (nine for the string, one for the feature), we had to split différenti (as in différentiation) by placing diff on the prefix table and érenti on the root table.

As shown on Figure 2 (page 209), a word may go through the root table several times depending on whether a prefix is found.

Perhaps, the most important thing we have learned while working on this project is that the efficient use of computers in linguistics requires that we look at problems from the computer's point of view. Computers are vastly superior to the human mind in some respects, but they are vastly inferior in some other respects. Our mind arrives at decisions in a "human" way; computers can do so only in a "machine" way and, although we are the creators of the machine, it is not easy to abandon our mental processes and create a "reasoning" process which utilizes to the maximum the resources of the computer. In our program, the lexicon, the root table, the prefix table, the ending table, the transcription rules, the transformation rules, etc. are all interdependent; they are meshed in such a way that a solution path can always be found for any problem. Présupposer, désolidariser with /s/ and présurer, désorganiser with /z/ are examples of words which the human mind and the computer must treat differently. In order to reach the right solution in the most efficient manner, computers may have to follow a series of steps which to the human mind seems absurd.

D. THE PREFIX TABLE

1. THE NEED FOR A PREFIX TABLE

Our program contains an algorithm which separates prefixes; for example, entretient is treated as entre + tient. This algorithm is needed to isolate the roots and words which do not follow regular transcription rules.

If we did not have that algorithm, the words:

anticognition	autocognition	extracognition
hypercognition	omnicognition	paracognition
précognition	pseudocognition	supracognition
ultracognition		

would have to be listed individually since, in the root cognit, the string gn has an irregular pronunciation. With our algorithm, it is sufficient to enter the root cognit only once.

Likewise, the words:

tient contient détient entretient retient

would have to be entered separately since, in all those verbs, the string en does not have its usual pronunciation. With our algorithm, it is sufficient to enter the exception word tient only once.

Another reason for having a prefix table is that some prefixes are so heavily used (e.g., anti, multi, radio, super) that it is impossible to list all the combinations of prefix + word from our lexicon which are now in use or which may be used in the future. For example, we have never met the word radiozinc nor is it listed in the Petit Robert (1979), but since radiocobalt is now in use, it is not unreasonable to believe that radiozinc may be used in the future or that it has already been used. In any case, our program is ready for it since we have radio on the prefix list and zinc in our lexicon.

2. SHOULD THE PREFIX TABLE INCLUDE ALL PREFIXES?

French uses about 300 prefixes. Some of those are so frequent (e.g., anti, auto, extra, re) that they must be included. Others are included in order to solve transcription problems; if hyper, inter, super were not included, the words hyperémotivité, interosseux, superovarié, etc. would require special rules to transcribe the letter e with /ɛ/ rather than with /ə/ (which is the usual transcription when e occurs before a syllabic cut).

But we do not believe it is necessary to include all the prefixes in our algorithm. Some (e.g., acro, hémo, litho) occur or are likely to occur only with roots which are not troublesome. Some occur in so few problem words (e.g., liposoluble) that it is easier to add those words to the lexicon. Finally, including some prefixes like épi, péri, pro would create more problems than it would solve; the lexicon and/or the root table offer better solutions. Our list of prefixes is in Appendix X (page 191).

3. UNACCEPTABLE PREFIX SEPARATIONS

We had to decide whether our program should be able to distinguish true prefixes (e.g., antivol, parapluie, restructurer) from non-prefixes (e.g., antilope, paradis, rescription), whether it should recognize the cases in which a particular string is no longer perceived as a prefix (e.g., interrogation versus interrègne), and whether it should always be able to

identify the real prefix (téléo in téléologie, but télé in téléopérateur). Writing a semantically-oriented program which could separate only true prefixes would require large amounts of time, money, and computer memory. On the other hand, allowing the computer to make wrong prefix separations (e.g., anti + lope) does not usually complicate our transcription rules or the analysis described on pages 52-82. Therefore, we decided not to attempt to separate only true prefixes and, in this report, the terms "prefix" and "prefixed word" do not necessarily mean that the word being discussed is a combination of a true prefix and a root. However, this decision requires that we catch all the prefix separations which would cause problems. We have three ways of preventing prefix separation: a) set a condition for prefix separation in the prefix table (page 191), b) use the root table, c) use the lexicon; if the lexicon is used, all the forms of the word must be listed (e.g., rescapé, rescapée, rescapés, rescapées; autour, autours) since the program goes through the prefix table before it goes through the table of endings and before final s is removed.

The prefix separations which must be prevented are as follows:

a. These words and their derivatives:

paraître	métairie	transept	métaux
parangon	microhms	macroure	reine
automne	monôme	métayer	autour

must not be interpreted as:

para + ître	méta + irie	trans + ept	méta + ux
para + ngon	micro + hms	macro + ure	re + ine
auto + mne	mono + ^me	méta + yer	auto + ur

b. We have to prevent prefix separation when pré is followed by s and when inter is followed by r since we cannot predict whether s is to be pronounced /s/ (e.g., préséance) or /z/ (e.g., présence) and whether we should have one /r/ (e.g., interrogation) or two (e.g., interrègne). The string dés is to be separated as dé in some cases (e.g., dé + ségrégation) and as dés in other cases (e.g., dés + engagement).

c. The string re can be separated when it is pronounced /rə/:

recharger	rescolariser	restructurer

but it must not be separated when it is pronounced /re/ or /rɛ/:

rectangle	respect	restaurant	ressayer	rescinder
ressuyons	reddition	renne	requiem	rescapé
rescousse	rescription	rescisoire	resta	restringent

or when it combines with m or n to form a nasal vowel:

rentrer	renivrer	renherber	rembourser

d. Separation must not be allowed if the remainder after separation is

the same as a word in the lexicon with a different pronunciation:

Lexicon words:	Separation cannot be allowed in:
ciment	déciment
cul	recul(s)
dent	extradent, oxydent, paradent
der	extrader, oxyder, parader
des	oxydes, parades, extrades, indes, ennéades, électrodes
fer	parafer
latent	relatent
mes	idiomes, minimes
sent	similisent
ses	hypnoses, orthoses, psychoses, similises
talent	détalent
ter	conter
tes	contes, idiotes, biotes
vent	archivent
ver	archiver
vient	dévient

The same problem could arise with a sequence of prefixes followed by a lexicon word (e.g., *antimicrotes), but we have not met any such word.

e. It must be prevented when the pronunciation of the remainder is not the same with every prefix. Sometimes the remainder is not an actual word; in the following case, only incident needs to be in the lexicon:

in + cident (/ɛ̃sidɑ̃/) dé + cident (/desid/)

Sometimes the remainder is an actual word; scient and valent keep their pronunciation in rescient and dévalent, but have a different pronunciation in conscient, omniscient, monovalent, plurivalent, and polyvalent; these last five words must be placed in the lexicon.

f. Separation must be prevented when the "prefix" requires a pronunciation different from the one used by our prefix table. Examples:

archi + atre con + nivence em + menthal

g. Separation is prevented when a semivowel, rather than a vowel, is used (words which, like uniaxe, require a vowel are in the lexicon):

antienne omniums oxyure polyandre polyèdre unionisme

h. What do we do if the remainder is such that it cannot be transcribed correctly by our transcription rules (e.g., détritus)? If the remainder has only one form (e.g., détritus, consensus, préfix, antillais do not have plural forms) and if the remainder is not used with other prefixes (e.g., retritus, contritus, etc. do not exist), it is preferable to place such words in the lexicon.

If the word has a plural in s (e.g., déficit/déficits), the lexicon

must contain either both forms of the complete word or the singular form of the remainder; we have adopted this second option. Examples:

(in)stinct	(con)cept	(hypo)scenium
(in)eptie	(idio)tie	(anti)llaise

If the plural is al/aux (e.g., quadriennal/quadriennaux), both the singular and the plural must be in the lexicon.

If the remainder is a word which occurs with one or several prefixes (e.g., détient, contient, entretient, retient), only this word (tient) is placed in the lexicon.

i. We do not allow prefix separation when the prefix is followed by only one or two characters since, without any gains, it would unduly complicate the transcription of words such as:

autos	as	auto + s	idiot	as	idio + t
micros	as	micro + s	idiots	as	idio + ts
paras	as	para + s	idiote	as	idio + te
radios	as	radio + s	idiome	as	idio + me
métal	as	méta + l	hypnose	as	hypno + se
néon	as	néo + n	psychose	as	psycho + se
électron	as	électro + n	postes	as	post + es
poste	as	post + e	métaux	as	méta + ux

Not separating prefixes followed by only one or two characters facilitates our work in many cases, but it does present a few problems:

--transit/transits: both forms must appear in the lexicon since the prefix is separated only in the plural.
--ennéade/ennéades: in the singular, the prefix is not separated and our rules would not transcribe initial enn correctly; in the plural, the prefix is separated, but the remainder des would be given the wrong pronunciation.
--resu/resus/resue/resues: only resu needs to be on the list.
--transi/transis/transie/transies: transi is in the lexicon and also serves for transie and transis; in transies, the prefix is separated and our rules transcribe the word correctly.

[Our lexicon contains two exception words with two-characters (us, ut) and three ambiguous words with two characters (as, Eu, os). The Petit Robert (1979) does not list any words combining a prefix + any of those five words. In any case, such words would probably use hyphens: pseudo-as, simili-os, etc. The past participle of ravoir (*reu) is not used.]

j. Our lexicon and our table of endings assign grammar categories. We must foresee the cases in which a word found in the lexicon (e.g., dis) or an ending from the table of endings (e.g., ending ère which assigns the category "noun") would assign a wrong category to prefixed forms (e.g., paradis marked as a verb or incinère marked as a noun). If the unprefixed form does not exist (e.g., cinère from incinère), we prefer to place in the lexicon the form without the prefix (e.g., cinère) —which will catch

the form without s and the form with s (if it exists)— rather than to have to list two complete words (e.g., incinère, incinères). If the unprefixed form exists (e.g., dé/lit), our program provides two ways to solve the problem. One way is to use the feature "npref" (bit 64) to indicate which grammar category is to be used if a prefix has been separated; for example, boucher appears as follows in the lexicon:

boucher buʃe inf,npref,nninf

This code tells the computer to use the feature "nninf" if boucher has been used without a prefix and to use the feature "inf" if a prefix has been separated (e.g., déboucher, emboucher, reboucher). This method saves computer memory since we need only one line of code instead of several, but we must be sure that all the prefixed words belong to the same category; for example, we could not use it for lit since délit is a noun and relit is a verb; likewise, if someone made up the word superboucher, our program would mark it as an infinitive, which is obviously wrong. The other way uses more space, but is safer; it consists in placing the prefixed words in the lexicon. In some cases, we need only the form without s (e.g., consort) since consorts and sorts are both nouns; in other cases, we need only the form with s (e.g., démets) since démet and met are both verbs; in other cases, we need both forms (e.g., décent, décents) since cent/s are numbers while décent/s are adjectives; and in other cases, we need four forms (e.g., dépoli/e/s) since poli and polis can be nouns.

4. COMBINATIONS OF PREFIXES AND AMBIGUOUS SPELLINGS

When an ambiguous spelling (e.g., portions, fier) is combined with a prefix, two possibilities occur:

a. In a few cases, the spelling remains ambiguous (e.g., infections/désinfections, affections/désaffections, contractions/décontractions, inventions/réinventions); only the words which use a prefix not on our list (e.g., ré) need to be listed.

b. In the other cases, the prefixed spelling is not ambiguous (e.g., parent/préparent, portions/emportions, fier/confier). We must separate the words which use a prefix listed in our table from those which do not.

- The prefixes which are not on our prefix table are the least usual and we find that practically all the words combining one of those prefixes with an ambiguous spelling are transcribed correctly by our rules:

comparent	importions	mécontentions	réparent	composter
remportions	exportions	comportions	encarter	enrhume

Words which our rules could not transcribe correctly (e.g., airbus, bibliobus, disconvient, circonvient, contresens) are in the lexicon.

- Most of the prefixes on our prefix table are heavily used and there is no limit to the combinations which can be made; for example, antiviolent, archicontent, autoconvergent are not in the Petit Robert (1979) although they are immediately understandable. Therefore we perform

the disambiguation analysis when our program finds a word composed of one of our prefixes and an ambiguous spelling except for the words listed below which have been placed in the lexicon to prevent prefix separation:

antibois	antiferment	autobus	confier
déchut	décorner	défier	démentions
dénotions	déparent	déplacer	déportions
dérations	désemparent	emparent	emportions
empressent	entrelacs	hypersécrétions	hyposécrétions
microbus	minibus	omnibus	polyester
prénotions	préparent	radioreporter	referment
relis	relut	renotions	replacer
reportions	revis	téléreporter	transparent
transportions	unifier		

5. PREFIXES OTHER THAN THOSE ON OUR LIST

a. Our prefix list does not contain all the prefixes used in French because there are prefixes whose inclusion on the list would present more drawbacks than advantages, but we have to be careful to place in the lexicon or the root table the words which begin with one of these prefixes and which would not go through the transcription rules correctly (e.g., cosujet, cosinus, prosimiens, girasol, homosexuel, réexamen).

b. In a few rare cases, the fact that a prefix is not on our list (e.g., ethno) may prevent the proper functioning of the prefix table (e.g., the prefix psycho in ethnopsychologie would not be found) and the proper functioning of the root table (e.g., the root psychia in ethnopsychiatrie would not be found). To solve these problems we use the transcription rules or the lexicon.

E. THE TABLE OF ENDINGS

1. General comments:

The ending search (see Appendix XI, page 193) is one of the most important steps in our program because:

a. By rewriting some of the endings (e.g., ments is changed to man, ations is changed to assion), we greatly simplify the transcription rules.

b. In French, word endings provide a convenient means of assigning linguistic features to words. A considerable amount of time has been spent on discovering which endings provide the most information. [The various ways of assigning these features are discussed in Appendix VII, page 182.]

The words which are exceptions to the classifications set up by the table of endings are in Group One of the lexicon (beginning on page 87).

For example, this group contains:

- gemment because it is the only French word ending with emment

which is not an adverb.

- acquiescent, immiscent because they are the only French words ending with scent which are not nouns.

The search stops as soon as a match is found; for example, the ending of réclamations is matched by the entry ations and the program does not go on to examine the other table entries. It is not necessary that the ending be modified (e.g., final stion remains stion): it is sufficient that it be found for the linguistic features to be properly assigned.

As shown on Figure 2 (page 209), the words which end with the letter s go through the ending search twice. It is, therefore, extremely important —whenever words ending with s do not have the same pronunciation or do not belong to the same grammar category as the words without s— to include both forms on the table of endings: ents and ent, ais and ai, ers and er, etc.

2. The following comments outline the process we used to choose and classify the 95 endings on our table:

a. ents and ent endings: Words ending with ents and ent constitute about 6% of the words in a narrative text. Since final s is removed before we reach the transcription rules, we must ensure that no words ending with ents will accidentally be treated as an ent verb.

The easiest way to do this is to change all the occurrences of ents to an. Since ents can occur after g and c, we must have separate statements for gents and cents to avoid the transcriptions /gā/ and /kā/. We are replacing gents with jan in order to keep the sound /ʒ/. We are replacing cents with çan because san would yield /zā/ after a vowel and ssan would give /ssā/ after s (thus, effervescents would be transcribed /efɛrvɛssā/ instead of /efɛrvesā/).

Since a /z/ of liaison is always acceptable between a plural noun ending with ents and an adjective beginning with a true vowel or semivowel or between a plural adjective ending with ents and a noun beginning with a true vowel or semivowel, there is no need —in view of the objectives of this program— to separate adjectives ending with ents from nouns ending with ents. Thus, we can have the following statements:

gents	jan	adj,m,pl	(no exceptions in lexicon)
cents	çan	adj,m,pl	(one exception in lexicon)
ents	an	nn,m,pl	(no exceptions in lexicon)

There are well over 1500 adverbs, nouns, and adjectives ending with ent and thousands of verbs ending with ent (e.g., chantent, chantaient, chantèrent, chanteraient, chantassent, viennent, venaient, vinrent, viendraient, vinssent; sent, ment, repent; devient, tient, vient). The 19 categories we have established have only 314 exceptions (which we have added to the lexicon) and all these categories assign an explicit[14] linguistic feature except for one which assigns the feature "nnadv".

The ending emment is changed to eaman (with an extra e) so that words like innocemment and intelligemment will be transcribed with /s/ and /ʒ/.

The ending scent is changed to ssan rather than çan to insure that the letter e in words like adolescent will be transcribed with /e/ and not /ə/ (as it would if it were followed by only one consonant).

We must ensure that removing the final nt of a verb does not create a word which appears in the lexicon with a pronunciation other than that of the verb. The only case we have found is veule/veulent.

Removing the final nt of a verb may create a word which appears in the lexicon with the same pronunciation, but with different linguistic features (e.g., the verb jalousent and the adjective jalouse): this does not matter since any explicit linguistic feature assignment made by the table of endings is never modified by the lexicon.

b. tion and tions endings: Since our program removes final s after the first pass through the table of endings, we must ensure that all the problems posed by final tions are solved during the first pass through the table of endings.

In this part of the program, we attempt to find categories which yield the maximum amount of linguistic information with the minimum number of exceptions. A study of words ending with tions shows that all words ending with ations (except 20), with ctions (except 5), and with ptions (except 9) are feminine plural nouns pronounced with /s/ and that the remaining words ending with tions (except 141) are verbs pronounced with /t/.

Thirty-nine words ending with tions which can be nouns (pronounced with /s/) or verbs (pronounced with /t/) are on the list of ambiguous spellings (e.g., adoptions, rations).

We need only three categories for the singular; all words ending with stion are singular nouns, all words ending with xtion are feminine singular nouns, and all remaining words ending with tion (except two) are feminine singular nouns.

F. REMOVING FINAL s

We remove final s since, in the vast majority of cases, it is not pronounced.

The only potential problem occurs with nouns or adjectives like remous, that is words which always end with s and which have before the final s a string which appears on the table of endings (e.g., ou). Those words must be included in the lexicon so that they will be marked "ambiguous number" rather than "plural".

G. REWRITE FOR THE LETTER h

The next step in our computer program deals with the letter h.

1. h after c, p, or s:

The combinations ch, ph, and sh are not modified. Words such as déshabiller, transhumance are transcribed correctly since the prefixes dés and trans are separated before we reach this point in our program. Words such as bonshommes, gentilshommes and proper names such as Boishardy, Deshayes, Malesherbes must be in the lexicon.

2. Other occurrences of h:

a. Final h (e.g., fellah), h before consonant (e.g., dahlia), h after consonant (e.g., thé) are removed. [Words in which the removal of h presents a problem (e.g., mathieu) have been added to the lexicon.]

b. Intervocalic h is kept in order to facilitate the transcription of words such as envahir. It is dropped during the transformation phase of our program (page 52).

c. Initial h is kept; its transcription poses several problems which are examined in Appendix XII, page 194.

H. REWRITE FOR INTERVOCALIC s

Intervocalic s is changed to z. All words in which intervocalic s is pronounced /s/ (e.g., vivisection) go through the root table, the prefix table, or the lexicon.

I. SYLLABICATION

In order to transcribe the letter e accurately, we need to know whether it immediately precedes a syllabic cut. We use the following syllabication scheme:

1. We place the syllabic cut after the vowel if the vowel is followed:

a. by only one consonant:

che.vaux ne.veu pe.ti.te

b. by a consonant other than a liquid or nasal (l, r, m, n) + a liquid (l, r):

Le.blond che.vreau se.vrer

c. by ch, ph, or sh followed by a vowel or a liquid (l, r):

de.re.chef Le.chat De.phlot

2. In other cases, the syllabic cut is placed after the first consonant (ch, ph, and sh count for only one consonant; other letters h, except intervocalic h and initial h, have already been removed):

res.te	seg.men.te	scep.tre
ef.fort	en.nui	em.me.ner
en.le.vé	gen.re	en.gen.dré
er.re.ment	mer.lan	sel.let.te
ver.nis	ber.cé	ec.clé.sias.ti.que
tech.ni.que	mech.ta	ec.chy.mo.se

3. Our algorithm does not place a syllabic cut between vowels (e.g., gé.ant) since this information is never needed by our transcription rules.

4. We have a special algorithm for ex (e.g., examiner); see page 38.

5. The string egn occurs only in proper names (e.g., Regnard) and is treated separately (see page 38).

J. TRANSCRIPTION RULES

1. General comments:

a. We use the traditional phonemic analysis with 16 vowels and three semivowels; to the usual 17 consonants, we have added /ŋ/ for borrowings such as camping.

b. The transcription rules are only one of the links in the total program. In order to be solved, some problems must be treated at various spots in the program: root table, prefix table, table of endings, transcription rules, transformation rules, etc.

c. For the vowel phonemes /e/, /ø/, /o/, /ɛ/, /œ/, and /ɔ/, we have adopted the pronunciation recommended in the Petit Robert (1979).

d. The phonemes /a/ and /ɑ/ are discussed in Appendix XIII, page 196.

e. The phoneme /ɲ/ is discussed in Appendix XIV, page 199.

f. The problems posed by the use of oe instead of œ in bœuf, fœtus, etc. are discussed in Appendix XV, page 199.

g. The problems posed by the unstable vowel (spelled e, ai, and on) are discussed in Appendix XVI, page 201.

h. In 1976, the French Academy adopted some spelling reforms (see Grevisse (1980), page 151). For example, événement, réglementaire, aiguë, gageure may now be spelled évènement, règlementaire, aigüe, and gageüre. Our program gives the correct pronunciation for either form.

i. We transcribe double written consonants with one sound (e.g., /gramɛrø/ for grammaire) except for initial ill (e.g., /illizibløø/ for

illisible) and except for words like transsibérien, interrègne, surréalisme in which the coexistence of two separate strings is clearly felt.

j. Since the degree of assimilation which takes place in clusters such as /b + s/ (e.g., absolu), /d + s/ (e.g., médecin) varies with several factors (e.g., speed of delivery), we use the transcriptions /bs/, /ds/ rather than /ps/, /ts/.

k. In view of the operation of our program, a transcription rule is not necessarily true when taken in isolation; for example, our rule that final en is /ɑ̃/ is acceptable only because strings like éen, ien, yen have been treated in preceding rules. The statement that ai is /ɛ/ is acceptable only because the table of endings has changed final ai into ê and changed final ais, ait, and aient into ai.

l. At this point in our computer program, only the characters or strings of characters shown in the following 28 units are acceptable. Foreign characters (e.g., Czech ř) or slashes not already treated by our code cause the program to stop and a message is sent to the user.

2. The transcription rules look at each character or string of characters from left to right. For example, bouillon is treated as follows:

Since the word begins with the letter b, the program calls a portion of code named "unit b" and, starting from the top, scans each statement until it finds that a single letter b is transcribed /b/. The letter b is removed, the program moves to "unit o" where it finds that the five-character string ouill is transcribed /uj/. The string ouill is removed and the program stays in "unit o" since the next character is o. In "unit o", the program finds that final on is transcribed /ɔ̃/.

In each unit, it is important that the statements be listed in the correct order. For example, in "unit a", the statement about ault must precede the statement about au; if not, the statement about ault would never be found.

3. The transcription rules use these definitions:

V	= any vowel
S	= e, é, è, ê, ë, i, ï, î, y, ÿ
C	= any consonant
N	= consonants m or n
nN	= non-nasal consonant
NnN	= nasal consonant followed by non-nasal consonant
NC	= nasal consonant followed by any consonant
NNV	= two nasal consonants followed by any vowel
L	= consonants l or r
nN./	= non-nasal consonant followed by syllabic cut
/.C	= consonant preceded by syllabic cut

4. The transcription rules are as follows:

Strings which have been modified by the various tables are shown

in their modified forms (e.g., final er changed to é, root boycott changed to bwwkott, apostrophe rewritten as iy, intervocalic s changed to z).

unit: a

If:	Remove:	Transcribe:
a in final position	a	a
at in final position	at	a
ayeri	aye	ɛjə
aieri	aie	ɛ
aze in final position	aze	ɑzɸ
abre in final position	abre	ɑbrɸ
adre in final position	adre	ɑdrɸ
afle in final position	afle	ɑflɸ
avre in final position	avre	ɑvrɸ
ault in final position	ault	o
auld in final position	auld	o
au	au	o
amp in final position	amp	ɑ̃
anm	an	ɑ̃
ainm	ain	ɛ̃
a + NnN	a + N	ɑ̃
an in final position	an	ɑ̃
ai + NnN	ai + N	ɛ̃
ai + N in final position	ai + N	ɛ̃
aill	aill	aj
ail in final position	ail	aj
ai	ai	ɛ
ay in final position	ay	ɛ
ay followed by consonant	ay	ɛ
aye in final position	aye	ɛj
ay	ay	ɛj
a	a	a

unit: â

If:	Remove:	Transcribe:
âill	âill	ɑj
âme in final position	âme	amɸ
âte in final position	âte	atɸ
ât in final position	ât	a
â	â	ɑ

unit: b

It is preferable to transcribe final b as /b/ since the number of words in which it is pronounced is greater than the number of words in which it is not pronounced and all the more so since new formations or borrowings pronounce it (e.g., lob, rab, toubib).

If:	Remove:	Transcribe:
bbl	bbl	bl
bed + V, in initial position	bed	bəd

bel + V, in initial position	bel	bəl
ben + V, in initial position	ben	bən
bez + V, in initial position	bez	bəz
bluff	bluff	blœf
bl	bl	bl
bbr	bbr	br
br	br	br
basse in initial position	bass	bɑs
bapt	bapt	bat
bb	bb	b
bwwkott	bwwkott	bɔjkɔt
bataill in initial position	bataill	batɑj
bach in final position	bach	bak
berg in final position	berg	bɛrg
b	b	b

unit: c

If:	Remove:	Transcribe:
ccl	ccl	kl
class in initial position	class	klɑs
cl	cl	kl
ccr	ccr	kr
crawl	crawl	krol
cr	cr	kr
cch + V	cch	k
chlor	chlor	klɔr
chr	chr	kr
champl in initial position	champ	ʃɑ̃p
champ + C, in initial position	champ	ʃɑ̃
ch + C, in initial position	ch	ʃ
ch + C	ch	k
chevel + V	chevel	ʃəvl
chevr in initial position	chevr	ʃəvr
chondr	chondr	kɔ̃dr
ch	ch	ʃ
cc + S	cc	ks
ciy in initial + final position	ciy	s
c + S	c	s
cqu	cqu	k
courra	courr	kurr
courré	courré	kurre
courron	courron	kurrɔ̃
courrion	courrion	kurrjɔ̃
courrié	courrié	kurrje
ccueill	ccueill	kœj
cueill	cueill	kœj
ccueil in final position	ccueil	kœj
cueil in final position	cueil	kœj
ct in final position	ct	kt
ck	ck	k
cuill	cuill	kɥij

compt	compt	kɔ̃t
cc	cc	k
coel + V	coe	se
coen + V	coe	se
c	c	k

unit: ç

If:	Remove:	Transcribe:
ç	ç	s

unit: d

If:	Remove:	Transcribe:
ddr	ddr	dr
dr	dr	dr
diy in initial + final position	diy	d
d in final position	d	-
deven + V	deven	dəvn
degr in initial position	degr	dəgr
den + V, in initial position	den	dən
doss	doss	dos
dompt	dompt	dɔ̃t
dd	dd	d
damn	damn	dɑn
d	d	d

unit: e

If:	Remove:	Transcribe:
éen in final position	éen	eɛ̃
é	é	e
è	è	ɛ
ê	ê	ɛ
ë	ë	ɛ
e in final position	e	ø
et in final position	et	ɛ
ed in final position	ed	e
emp in final position	emp	ɑ̃
en in final position	en	ɑ̃
eau	eau	o
ei + N in final position	ei + N	ɛ̃
ei + NnN	ei + N	ɛ̃
eill	eill	ɛj
eil in final position	eil	ɛj
ei	ei	ɛ
euill	euill	œj
euil in final position	euil	œj
eu in initial position	eu	ø
eur	eu	œ
euf in final position	euf	œf
euv	euv	œv

eul	eu	œ
eubl	eubl	œbl
eu	eu	ø
ey in final position	ey	ɛ
eyeri	eye	ɛjə
ey + C	ey	ɛ
eye in final position	eye	ɛj
ey	ey	ɛj
e + V	e	-
egn	e	ɸ
eria	eri	ərj
erien in final position	erien	ərjɛ̃
erie	e	ə
erio	eri	ərj
eriu	eri	ərj
elia	eli	əlj
elien in final position	elien	əljɛ̃
elie	e	ə
elio	eli	əlj
eliu	eli	əlj
eri	eri	əri
eli	eli	əli
eni + V	e	ə
ensevel + V	ensevel	ɑ̃səvl
ex + V, in initial position	ex	ɛgz

The following twenty strings are listed below so that they will be given their correct pronunciation should they occur after a prefix not included in our table (e.g., coexistence, réexaminer or inside words such as sexagésime.

exact	exa	ɛgza
exagé	exagé	ɛgzaʒe
exal	exal	ɛgzal
exam	exam	ɛgzam
exaspé	exaspé	ɛgzaspe
exau	exau	ɛgzo
exéc	exéc	ɛgzek
exemp	exem	ɛgzɑ̃
exer	exer	ɛgzɛr
exib	exib	ɛgzib
exum	exum	ɛgzym
exig + S	exig	ɛgziʒ
exigu	exigu	ɛgzigy
exil	exil	ɛgzil
exist	exist	ɛgzist
exonér	exonér	ɛgzɔner
exor	exor	ɛgzɔr
exoti	exoti	ɛgzɔti
exub	exub	ɛgzyb
exul	exul	ɛgzyl
ex + V	ex	ɛks

enn in initial position	enn	ɑ̃n
etnopsych	etnopsych	ɛtnɔpsik
e + /.C, in initial position	e	e
e + /.C	e	ɸ
esc + S	e	e
esç (as in acquiesça)	e	e
eff	e	e
ell	e	e
emm	e	e
enn	e	e
ess	e	e
ett	e	e
e + NNV	e	ɛ
e + NnN	e + N	ɑ̃
e	e	ɛ

unit: f

If:	Remove:	Transcribe:
ffl	ffl	fl
fl	fl	fl
ffr	ffr	fr
fr	fr	fr
faiza in initial position	faiz	fəz
faizeu in initial position	faiz	fəz
fenouil in initial position	fen	fən
foss	foss	fos
faiz	faiz	fɸz
ff	ff	f
fabl in initial position	fabl	fɑbl
foet + V	foe	fe
fascis	fasci	faʃi
fasciz	fasci	faʃi
f	f	f

unit: g

It is preferable to transcribe final g (except ng) as /g/ since the number of words in which it is pronounced is greater than the number of words in which it is not pronounced.

If:	Remove:	Transcribe:
gross	gross	gros
ggr	ggr	gr
gr	gr	gr
ggl	ggl	gl
gl	gl	gl
g in final position	g	g
gueill	gueill	gœj
gueil in final position	gueil	gœj
guë in final position	guë	gy
guî + C	guî	gɥi
gü	gü	gy

guy + V	guy	gɥij
gu + V	gu	g
geü	geü	ʒy
geu	geu	ʒø
ggé	ggé	gʒe
ggè	ggè	gʒɛ
gge + C	gge	gʒɛ
géa	g	ʒ
géo	g	ʒ
gea	ge	ʒ
geo	ge	ʒ
gagn in initial position	gagn	gɑɲ
g + S	g	ʒ
gg + V	gg	g
g + V	g	g
gn	gn	ɲ
gt in final position	gt	-
gt + V	gt	t
g	g	g

unit: h

If:	Remove:	Transcribe:
habi in initial position	hab	ab
hagio in initial position	hagi	aʒj
halen in initial position	halen	aln
halèn in initial position	halèn	alɛn
halitu in initial position	hal	al
hallucl in initial position	hallucl	alysi
halog in initial position	hal	al
haloph in initial position	haloph	alɔf
haltê in initial position	haltê	alte
hameç in initial position	hameç	ams
haplo in initial position	hapl	apl
harmo in initial position	harm	arm
ha in initial position	h	'
heurt in initial position	h	'
herb in initial position	herb	ɛrb
héral in initial position	héral	eral
hercu in initial position	herc	ɛrk
hercy in initial position	hercy	ɛrsi
hérê in initial position	hérê	ere
hérit in initial position	hérit	erit
herm in initial position	herm	ɛrm
hertz in initial position	hertz	ɛrts
héroï in initial position	héroï	erɔi
herpé in initial position	herpé	ɛrpe
her in initial position	h	'
hebdo in initial position	hebd	ɛbd
héb in initial position	héb	eb
hect in initial position	hect	ɛkt
héd in initial position	héd	ed
hég in initial position	hég	eʒ

hẽli in initial position	hẽl	el
hel + C in initial position	he	ɛ
hém in initial position	hém	em
hend in initial position	hend	ɛd
hépa in initial position	hép	ep
hepta in initial position	hept	ɛpt
hẽz in initial position	hẽz	ez
hẽt in initial position	hẽt	et
heur in initial position	heu	œ
hex + V in initial position	hex	ɛgz
he in initial position	h	ʻ
hiata in initial position	hiat	jat
hiber in initial position	hiber	ibɛr
hiém in initial position	hiém	jem
hiérat in initial position	hiérat	jerat
hiéro in initial position	hiér	jer
hilar in initial position	hilar	ilar
hilo in initial position	hil	il
himalay in initial position	himala	imala
hin in initial position	h	-
hipp in initial position	hipp	ip
hir in initial position	hir	ir
hisp in initial position	hisp	isp
hist in initial position	hist	ist
hitl in initial position	hitl	itl
hiv in initial position	hiv	iv
hi in initial position	h	ʻ
hoir in initial position	hoir	war
holo in initial position	hol	ɔl
homẽ in initial position	homẽ	ɔme
homi in initial position	hom	ɔm
homm in initial position	homm	ɔm
homo in initial position	hom	ɔm
honnẽ in initial position	honnẽ	ɔnɛ
hono in initial position	hon	ɔn
hõpi in initial position	hõp	ɔp
hordé in initial position	hordé	ɔrde
horiz in initial position	horiz	ɔriz
horl in initial position	horl	ɔrl
hormo in initial position	horm	ɔrm
horo in initial position	hor	ɔr
horr in initial position	horr	ɔr
hort in initial position	hort	ɔrt
hos in initial position	ho	ɔ
hoz in initial position	ho	ɔ
hõt in initial position	hõt	ot
ho in initial position	h	ʻ
huil in initial position	huil	ɥil
huis in initial position	hui	ɥi
huît in initial position	hui	ɥi
humai in initial position	hum	ym
human in initial position	hum	ym
humb in initial position	humb	œb

humect in initial position	humect	ymɛkt
humér in initial position	humér	ymer
humi in initial position	hum	ym
humor in initial position	humor	ymɔr
hu in initial position	h	ʻ
hy in initial position	h	-

unit: i

If:	Remove:	Transcribe:	
ien in final position	ien	jɛ̃	
in + nN	in	ɛ̃	
i	i	i	
i + NC	i + N	ɛ̃	
i	i	i	
i in final position	i	i	
it in final position	it	i	
ien in final position	ien	jɛ̃	
in in final position	in	ɛ̃	
ieri	ie	i	
iendr	iendr	jɛ̃dr	
illaire in final position	illaire	ilɛrø	
illon in initial position	ill	ij	(ex: post + illon)
ill in initial position	ill	ill	
ill	ill	ij	
inm	in	ɛ̃	
i + NnN	i + N	ɛ̃	
ign in final position	ign	iŋ	
i	i	i	

unit: j

If:	Remove:	Transcribe:
jiy in initial + final position	jiy	ʒ
jeũn	jeũn	ʒøn
jeun + V	jeun	ʒœn
jj	jj	ʒ
j	j	ʒ

unit: k

If:	Remove:	Transcribe:
kiy in initial + final position	kiy	k
kreml	kreml	krɛml
kk	kk	k
k	k	k

unit: l

If:	Remove:	Transcribe:
ll	ll	l
liy in initial + final position	liy	l
lt in final position	lt	lt
ld in final position	ld	ld
lockout in initial position	lockout	lɔkawt
levr in initial position	levr	lɔvr
l	l	l

unit: m

If:	Remove:	Transcribe:
mm	mm	m
miy in initial + final position	miy	m
mourr	mourr	murr
meul	meul	møl
montr	montr	mɔ̃tr
mont + C, in initial position	mont	mɔ̃
mer + V, in initial position	mer	mar
m	m	m

unit: n

If:	Remove:	Transcribe:
nn	nn	n
niy in initial + final position	niy	n
neur in initial position	neur	nør
n	n	n

unit: o

If:	Remove:	Transcribe:
ô	ô	o
o in final position	o	o
ourg in final position	ourg	ur
ould in final position	ould	u
oult in final position	oult	u
od in final position	od	o
ot in final position	ot	o
ou in final position	ou	u
oz	o	o
oin in final position	oin	wɛ̃
oueri	oue	u
oieri	oie	wa
oi + NnN	oi + N	wɛ̃
oi	oi	wa
om in final position	om	ɔm
omb in final position	omb	ɔ̃
omp in final position	omp	ɔ̃
on in final position	on	ɔ̃

onm	on	ɔ̃
o + NnN	o + N	ɔ̃
ouill	ouill	uj
ouil in final position	ouil	uj
oui + V	oui	uj
ou	ou	u
oyen in final position	oyen	wajɛ̃
oy in final position	oy	wa
oy + C	oy	wa
oy	oy	waj
ossion in final position	ossion	osjɔ̃
orchestr	orchestr	ɔrkɛstr
och in final position	och	ɔk
ota in initial position	ot	ot
ote in initial position	ot	ot
oti in initial position	ot	ot
oeill	oeill	œj
oeil in final position	oeil	œj
oeu in final position	oeu	ø
oeud in final position	oeud	ø
oeu	oeu	œ
oezoph + V	oezoph	ezɔf
oecum + V	oecu	eky
oed + V	oe	e
oen + V	oe	e
oestr + V	oestr	ɛstr
oe	o	ɔ
o	o	ɔ

unit: p

It is preferable to transcribe final p after vowel as /p/ since the number of words in which it is pronounced is greater than the number in which it is not pronounced and all the more so since new formations or borrowings pronounce it (e.g., beaup, bop, clip, drop, slip, stup).

If:	Remove:	Transcribe:
ppl	ppl	pl
peupl	peupl	pœpl
pl	pl	pl
ppr	ppr	pr
pr	pr	pr
ph	ph	f
passe	pass	pas
payz	payz	peiz
pp + V	pp	p
pelo in initial position	pel	pɛl
pelu in initial position	pel	pɛl
pel + V, in initial position	pel	pəl
pen + V, in initial position	pen	pən
pez + V, in initial position	pez	pəz
poec + V	poe	pe

p + V	p	p
pt in final position	pt	-
pp	pp	p
p	p	p

unit: q

If:	Remove:	Transcribe:
qū	qū	ky
querra	querr	kɛrr
querré in final position	querré	kɛrre
querron in final position	querron	kɛrrɜ̃
querrion in final position	querrion	kɛrrjɜ̃
querrié in final position	querrié	kɛrrje
quen + V, in initial position	quen	kən
quer + V, in initial position	quer	kɛr
qu	qu	k
qq	qq	k
q	q	k

unit: r

If:	Remove:	Transcribe:
ratio	rati	rasj
rr	rr	r
regard	regar	rǿgar
regret	re	rǿ
renon	re	rǿ
retour	retour	rǿtur
recev	recev	rǿsəv
repo	re	rǿ
ressembl	ressembl	rǿsɑ̃bl
ressemel	ressem	rǿsəm
restab	restab	rɛstab
restat	restat	rɛstat
raill in initial position	raill	raj
rumo in initial position	rum	rɔm
ruma in initial position	rum	rɔm
rumi in initial position	rum	rɔm
rumer in initial position	rum	rɔm
racle in final position	racle	rakl̃ǿ
r	r	r

unit: s

If:	Remove:	Transcribe:
sabl in initial position	sabl	sɑbl
sev in initial position	sev	sav
siy in initial + final position	siy	s
surre in initial position	surr	syrr
s + V	s	s
ss	ss	s
sc + S	sc	s

st	st	st
sph	sph	sf
sp	sp	sp
sl	sl	sl
shampoo	shampoo	ʃɑ̃pu
sh	sh	ʃ
schizo	schiz	skiz
sch	sch	ʃ
squa in initial position	squa	skwa
sculpt	sculpt	skylt
sç	sç	s
swingu	swin	swiŋ
s	s	s

unit: t

If:	Remove:	Transcribe:
t in final position	t	-
tiy in initial + final position	tiy	t
tiair	tiair	sjɛr
tial	tial	sjal
tiau	tiau	sjo
tieu	tieu	sjø
tiell	tiell	sjɛl
tiel	tiel	sjɛl
tion in final position	tion	tjɔ̃
tion	ti	sj
trust in initial position	trust	trœst
tasse	tass	tɑs
taill in initial position	taill	tɑj
tt	tt	t
tz	tz	ts
t	t	t

unit: u

If:	Remove:	Transcribe:
u in final position	u	y
uin in final position	uin	ɥɛ̃
uin + nN	uin	ɥɛ̃
un in final position	un	œ̃
ueri	ue	y
uieri	uie	ɥi
u + NnN	u + N	œ̃
um in final position	um	ɔm
uie	ui	ɥi
ui + V	ui	yj
ui	ui	ɥi
uy in final position	uy	ɥi
uy + C	uy	ɥi
uy	uy	ɥij
ulul in initial position	ulul	ˊylyl
u	u	y

unit: v

If:	Remove:	Transcribe:
vill in initial position	vill	vil
ville in final position	ville	vilø
viru in final position	viru	virys
veugl	veugl	vœgl
ved + V, in initial position	ved	vəd
vv	vv	v
v	v	v

unit: w

If:	Remove:	Transcribe:
ww	ww	v
wxz	wxz	gn
wzx	wzx	ñ
w	w	v

unit: x

If:	Remove:	Transcribe:
x in initial position	x	ks
xc + S	xc	ks
xc	xc	ksk
xt	xt	kst
xx in final position	xx	s
xx	xx	ks
x	x	ks

unit: y

If:	Remove:	Transcribe:
y + N in final position	y + N	ɛ̃
y + NnN	y + N	ɛ̃
yen in final position	yen	jɛ̃
yé in final position	yé	je
ye in final position	ye	i
y	y	i

unit: z

If:	Remove:	Transcribe:
zz	zz	z
z	z	z

5. After the word transcription is completed, the prefix(es), if any, is/are reattached.

K. TRANSFORMATION RULES

Deciding whether to use /i/ or /j/, /y/ or /ɥ/, /u/ or /w/, /e/ or /ɛ/, /ø/ or /œ/, /o/ or /ɔ/, /a/ or /ɑ/ is difficult because:

1. Native speakers of French do not always agree.

2. In the case of /e/ and /ɛ/, /ø/ and /œ/, /o/ and /ɔ/, intermediate timbres are often used (especially in unstressed position). Like most French dictionaries, we do not attempt to represent those intermediate timbres and, in general, we have the same transcriptions as those found in the Petit Robert (1979) (e.g., échelon is transcribed as /eʃlɔ̃/, but échevin is transcribed as /ɛʃvɛ̃/.

We have been able to take care of the problems posed by /o/, /ɔ/, /ø/, and /œ/ through the root table and the transcription rules. The problems posed by /a/, /ɑ/ are analyzed on pages 196-199. For /i/, /j/, /y/, /ɥ/, /u/, /w/, /e/, and /ɛ/, we have had to write a set of transformation rules.

The following transformation rules do not apply to the words found in the lexicon. Only rules 7 through 9 apply to the prefixes (e.g., /prevønir/ becomes /prɛvønir/); in other cases, the prefixes remain unchanged (e.g., /ãtiãʒɛ̃/). The rules must be applied in the order given below.

Rule 1:

1. If /ø/ occurs after a vowel, /ø/ is removed. Examples:

/tyøre/	becomes	/tyre/
/apɥiøra/	becomes	/apɥira/

-- Our transcription rules transcribe the letter e as /ə/ when it follows a consonant and precedes li, ni, or ri + a vowel (e.g., geliez, Richelieu, ameniez, mangerions). The letter e is not transcribed when it follows a vowel and precedes ri + vowel (e.g., tuerions, étudieriez, louerions); there are no examples of vowel + e + li or ni + vowel.

Rule 2:

2. If /y/ occurs before a vowel, /y/ is changed to /ɥ/:

/ˊya/	becomes	/ˊɥa/
/dyɛl/	becomes	/dɥɛl/
/sinyø/	becomes	/sinɥø/
/nyãsø/	becomes	/nɥãsø/
/ebɛrlye/	becomes	/ebɛrlɥe/
/afɛktyø/	becomes	/afɛktɥø/
/sɛksyɛl/	becomes	/sɛksɥɛl/
/vɛrtyø/	becomes	/vɛrtɥø/
/tymyltyø/	becomes	/tymyltɥø/

except when /y/ occurs after a non-liquid + liquid cluster:

/flyɔr/	remains	/flyɔr/
/tryã/	remains	/tryã/
/mɔ̃stryø/	remains	/mɔ̃stryø/

-- The spelling ui + vowel other than non-accented e (e.g., tuions)

is transcribed correctly with /yj/. The sequence non-liquid + liquid + ui (e.g., fruit, pluie) is transcribed with /ųi/; the few exceptions (e.g., truisme) have been placed in the lexicon.

-- The word duumvir is in the lexicon.

Rule 3:

3. If /u/ occurs before a vowel, /u/ is changed to /w/:

/uɛst/	becomes	/wɛst/
/ʒuɜ̃/	becomes	/ʒwɜ̃/
/nuɸ/	becomes	/nwɸ/
/syrlue/	becomes	/syrlwe/
/ɛskuadɸ/	becomes	/ɛskwadɸ/
/marsuɛ̃/	becomes	/marswɛ̃/

except when /u/ occurs after a non-liquid + liquid cluster:

/klue/	remains	/klue/
/ebluir/	remains	/ebluir/
/ɑ̃kluyrɸ/	remains	/ɑ̃kluyrɸ/

-- The spelling oui + vowel other than final e (e.g., secouiez) is transcribed correctly by our transcription rules with /uj/.

-- The word nouure is in the lexicon.

Rules 4-6:

Our transcription rules transcribe correctly the words where the graphemic cluster ill is followed by a vowel other than i: habillons, bâilla, cueillez, barbouillèrent, etc., but the problems posed by words such as oublions are such that we have adopted a two-step scheme: first transcription rules, and then a set of transformation rules.

The task is made all the more difficult since native speakers of French do not agree on the pronunciation of words such as habillions; some pronounce it as the present (habillons), others make a difference. The scheme used below is acceptable in careful speech and helps distinguish tenses (e.g., habillons vs. habillions, oublions vs. oubliions).

4. If /ii/ occurs in the sequence non-liquid + liquid + /ii/ + vowel, /ii/ is changed to /ijj/. Examples:

/ubliiɜ̃/	becomes	/ublijjɜ̃/
/kriie/	becomes	/krijje/
/striiɜ̃/	becomes	/strijjɜ̃/

In other cases, /ii/ followed by vowel becomes /ij/. Examples:

/siiɜ̃/	becomes	/sijɜ̃/
/etydiie/	becomes	/etydije/

/amnistiie/	becomes	/amnistije/
/ɛkspiiɜ̃/	becomes	/ɛkspijɜ̃/
/mariiɜ̃/	becomes	/marijɜ̃/
/marʃiie/	becomes	/marʃije/

However, in the case of the verb rire (but not sourire), we have opted for the pronunciations /rijjɜ̃/, /rijje/, and we have placed those words in the lexicon.

5. If /ji/ occurs before a vowel, /ji/ is changed to /jj/. Examples:

/abijiɜ̃/	becomes	/abijjɜ̃/
/revɛjiɜ̃/	becomes	/revɛjjɜ̃/
/travajie/	becomes	/travajje/
/ɑ̃vwajiɜ̃/	becomes	/ɑ̃vwajjɜ̃/
/barbujie/	becomes	/barbujje/
/mɔrdijiɜ̃/	becomes	/mɔrdijjɜ̃/

6. If /i/ occurs in the sequence non-liquid + liquid + /i/ + vowel, /i/ is changed to /ij/. Examples:

/fevrie/	becomes	/fevrije/
/pliɜ̃/	becomes	/plijɜ̃/
/sɥivrie/	becomes	/sɥivrije/
/sɑ̃blie/	becomes	/sɑ̃blije/

In other cases, /i/ followed by vowel becomes /j/. Examples:

/ʒɑ̃vie/	becomes	/ʒɑ̃vje/
/idio/	becomes	/idjo/
/aiœl/	becomes	/ajœl/
/iɔdø/	becomes	/jɔdø/
/pleiadø/	becomes	/plejadø/
/tarpeiɛnø/	becomes	/tarpejɛnø/
/asfiksia/	becomes	/asfiksja/
/kalɔmnie/	becomes	/kalɔmnje/
/pɛrlie/	becomes	/pɛrlje/
/sɔrtiɜ̃/	becomes	/sɔrtjɜ̃/
/murriɜ̃/	becomes	/murrjɜ̃/
/ʃɑ̃tərie/	becomes	/ʃɑ̃tərje/

However, we have placed these words in the lexicon with /ij/: ria, riais, riait, riaient, riant(e)(s), rieur(s), rieuse(s), riez, rions.

Rules 7-9:

With our transcription rules, é is always transcribed with /e/; the letter e, depending on its environment, receives several transcriptions, two of which are /e/ and /ɛ/. After each word has been transcribed, we must determine whether any of the /e/ should be changed to /ɛ/. In this set of transformations, we have adopted the system used by the authors of the Petit Robert (1979).

We use the following symbols:

C = any consonant sound
N = a nasal consonant: /m/, /n/, /ɲ/
L = a liquid consonant: /l/, /r/
X = any consonant except a liquid or a nasal: /t/, /s/, /b/, etc.
In the code, /w/ and /j/ must not be counted as consonants.

7. If /e/ occurs before one or two consonants followed by /ɸ/ in final position, /e/ is changed to /ɛ/. Examples:

/belɸ/	becomes	/bɛlɸ/
/saʒesɸ/	becomes	/saʒɛsɸ/
/refleksɸ/	becomes	/reflɛksɸ/

-- Before final consonant, our transcription rules always yield /ɛ/; no transformation is necessary.

8. This rule concerns /e/ occurring before consonant + /ɸ/ + consonant. Since /ɸ/ is normally dropped in this case (page 77), /e/ is actually preceding two consonants and its change into /ɛ/ depends on the nature of the consonants. The vowel /e/ is changed to /ɛ/ in these cases:

/e/ + N + /ɸ/ + L	crémerie, crénelé, régnera, vénerie, Fénelon
/e/ + N + /ɸ/ + N	événement, démener
/e/ + L + /ɸ/ + N	scellement, Barthélemy
/e/ + L + /ɸ/ + L	céleri, espérera
/e/ + L + /ɸ/ + X	pelletée, élever
/e/ + X + /ɸ/ + N	empiétement, abrégement, allégement, prévenir
/e/ + X + /ɸ/ + X	médecin, écheveau

/e/ remains /e/ when it occurs with the following clusters:

/e/ + X + /ɸ/ + L	cédera, éperon, échelon, betterave, décelé
/e/ + N + /ɸ/ + X	démesuré, sénevé

9. When /e/ precedes X + L + /ɸ/ (an /ɸ/ which will be changed to /ə/ on page 77), it normally remains /e/ (e.g., all the words using the root célébr), but there are erratic uses of /ɛ/. For these words, we use the lexicon (e.g., allégrement) or the root table (e.g., réglem).

We also keep /e/ in /e/ + /s/ + C (e.g., préscolaire, déstaliniser), but escalier, destination, etc. are correctly transcribed with /ɛ/.

Rule 10:

10. /ɛ/ is changed to /e/ when it occurs before one of these clusters (this is referred to as harmonisation vocalique):

/ɛ/ + C + /i/, /y/, or /e/: bêtise, laitue, têtu, entêté, éclairé

/ɛ/ + X + L + /i/, /y/, or /e/: aigri, salpêtrière

Rule 11:

11. Intervocalic h which had been kept to facilitate the transcription of words such as envahir, cahot, bahut is removed.

Rule 12:

12. Internal /ˇ/ (brought about by the reunification of a prefix with a hiatus word) is removed: /ɑ̃tiˇipi/ becomes /ɑ̃tiipi/.

VII. LINGUISTIC ANALYSIS OF THE INPUT STRING

At this point in our program, about 99% of the words and idioms have been transcribed into IPA symbols. The only words which have not been transcribed are those which have several pronunciations (e.g., fils).

About 65% of the transcribed words and idioms have been placed into an explicit grammar category (e.g., verb, noun); about 22% have been placed into an ambiguous grammar category (e.g., noun/adverb), and the rest have not been placed into any grammar category at all.

In this part of our program, we must:

a. Apply a set of rules designed to improve the linguistic analysis. This step is essential since the 65% level achieved so far is not usually sufficient to allow us to perform correctly the operations described in (b), (d), (e), (f), (g) below. It should be noted that the grammar categories we use are adapted to those specific needs (see Appendix VII, page 182, for a detailed discussion of this point). This step is not applied to the spellings which have been marked ambiguous (bit 71 set to 1).

b. Apply a set of rules to disambiguate the ambiguous spellings.

c. Repeat the first step if the second step was done and was successful since it may be possible to assign additional linguistic features after ambiguities have been resolved.

d. Divide strings which have more than fourteen syllables.

e. Apply rules to determine whether the pronunciation of numbers ending with /sɛ̆k/, /sis/, /ɥit/, /dis/ should be modified.

f. Add the consonants of liaison whenever necessary.

g. Decide which /ə/ must be changed to /ə/ and which must be dropped.

h. Present some test results.

A. RULES TO IMPROVE THE ASSIGNMENT OF LINGUISTIC FEATURES

One hundred and twelve bits have been reserved for each word of the

input string; the information held by each bit is described on pages 187-189. Since the explicit features are held by the first 18 bits, our task is to attempt to complete the analysis with each word having one (and only one) of those first 18 bits set to 1 or, if this fails, to reduce the linguistic range (e.g., change the linguistic feature from "adjnnant" which corresponds to bit 43 set to 1 to "adjnn" which corresponds to bit 19 set to 1).

To do so, our algorithm scans the input string and looks for words in which the sum of the values for bits 1 through 18 is equal to 0, indicating thereby that no explicit feature has been assigned. (We will use the letter "X" to refer to such words.)

1. First, the program considers the words which have been assigned an ambiguous grammar category (categories 19 through 51); it applies a first set of rules to each of these words. This first set of rules is repeated until a complete pass through the words with categories 19 through 51 fails to yield any further clarification.

We do not have sufficient space in this report to include all the rules of this first set; the rules used for category 19 are typical of those we use for the other categories:

Category 19 (adjective/noun) is changed to 4 (noun) if:

-- the a/n is preceded by a determiner or a number and is final of input string or is followed by a verb form, a pronoun, an adverb, or a word marked (neg)
-- the a/n is preceded by a determiner (or number or adverb) + an adjective and is final of input string or is followed by a verb form, a pronoun, an adverb, an adjective, or a word marked (neg)
-- the a/n is preceded by a determiner (or number or adverb) + an adjective + an adjective and is final of input string or is followed by a verb form, a pronoun, an adverb, an adjective, or a word marked (neg)
-- the a/n is final of input string and is preceded by an adjective

Category 19 is changed to 5 (adjective) if:

-- the a/n is preceded by a noun
-- the a/n is followed by a noun
-- the a/n is preceded by a verb form
-- the a/n is preceded by an adverb

2. After applying the first set of rules, the program considers the words which have not yet been assigned a grammar category from 1 through 18 and applies a second set of rules.

This second set of rules uses the lists shown on pages 204-207. The lists are numbered 77 through 112 and all the words contained in those 36 lists are in the lexicon. If a word from list 77 appears in the input string, bit 77 for that word is set to 1. A word may appear in several lists; for example, the word le appears in lists 78, 81, 88, 92, 94, 95 and those

six bits are set to 1 when the word le appears in an input string.

The program looks in the input string for sequences of words which fit the rules given below. For example, rule 4 means that if an X word is followed by a hyphen (bit 72 of X set to 1) and a word from list 99 and a word from list 79, then X must be an imperative. Rule 40 means that if the word ne is followed by a word from list 83 followed by a word from list 92 followed by an infinitive, then the sequence must be (neg)(adv)(di)(inf); this rule allows us, for example, to decide that pas is an adverb (not a noun) and that le is a pronoun (not a determiner).

After looking for each of the sequences described in the 113 rules, the program starts again at rule 1, and so on until all the words in the input string have been assigned a grammar category from 1 to 18 or until a complete pass through the 113 rules fails to identify any more words. At this point, we make another pass through the first set of rules since it may now be possible to assign an explicit grammar category to words which still have an ambiguous category.

The rules must be tested in the order given below. For example, rule 27 becomes true only after rules 14 through 26 have failed.

The notations X/inf, X/nn, etc. mean that the word has not yet been assigned a grammar category from 1 to 18 or that it has already been identified as an infinitive, a noun, etc.

Rule:	mark sequence as:
1. X (-)spr	v,spr signe-t-elle, dussé-je
2. X (-)lt77	v,spr caches-tu, fût-ce
3. X (-)lt78 (-)lt79	imp,di,di vends-le-moi, cache-les-y
4. X (-)lt99 lt79	imp,di,di pare-l'en
5. X (-)lt79 (-)lt79	imp,di,di vends-nous-en
6. X (-)lt78	imp,di vends-le
7. X (-)lt79	imp,di aide-moi
8. lt80 lt81 adv adv X/inf	prep,di,adv,adv,inf de le si bien conseiller
9. lt80 lt81 adv X/inf	prep,di,adv,inf de le bien conseiller
10. lt82 lt81 adv adv X/inf	prep,di,adv,adv,inf à le si bien conseiller
11. lt82 lt81 adv X/inf	prep,di,adv,inf à le bien conseiller
12. lt80 lt81 X/inf	prep,di,inf de le boucher

13. lt82 lt81 + X/inf	prep,di,inf à le boucher
14. neg lt83 lt84 lt78 X/inf	neg,adv,di,di,inf ne pas me le conseiller
15. neg lt83 lt84 lt85 X/inf	neg,adv,di,di,inf ne pas m'y tendre
16. neg lt83 lt84 lt86 X/inf	neg,adv,di,di,inf ne pas m'en devoir
17. neg lt83 lt78 lt87 X/inf	neg,adv,di,di,inf ne pas le lui conseiller
18. neg lt83 lt78 lt85 X/inf	neg,adv,di,di,inf ne pas les y noyer
19. neg lt83 lt78 lt86 X/inf	neg,adv,di,di,inf ne pas l'en pêcher
20. neg lt83 lt87 lt86 X/inf	neg,adv,di,di,inf ne pas leur en devoir
21. neg lt83 lt85 lt86 X/inf	neg,adv,di,di,inf ne pas y en avoir
22. neg lt83 adv adv lt88 X/inf	neg,adv,adv,adv,di,inf ne pas très bien le conseiller
23. neg lt83 adv lt88 X/inf	neg,adv,adv,di,inf ne pas bien le conseiller
24. neg lt83 lt88 X/inf	neg,adv,di,inf ne pas le conseiller
25. neg lt83 adv adv X/inf	neg,adv,adv,adv,inf ne pas très bien conseiller
26. neg lt83 adv X/inf	neg,adv,adv,inf ne pas bien conseiller
27. neg lt83 X/inf	neg,adv,inf ne pas conseiller
28. lt89 lt84 lt78 X/inf	prep,di,di,inf pour me le conseiller
29. lt89 lt84 lt85 X/inf	prep,di,di,inf pour m'y noyer
30. lt89 lt84 lt86 X/inf	prep,di,di,inf pour m'en devoir
31. lt89 lt78 lt87 X/inf	prep,di,di,inf pour le lui conseiller
32. lt89 lt78 lt85 X/inf	prep,di,di,inf pour les y noyer
33. lt89 lt78 lt86 X/inf	prep,di,di,inf pour l'en pêcher
34. lt89 lt87 lt86 X/inf	prep,di,di,inf pour leur en devoir
35. lt89 lt85 lt86 X/inf	prep,di,di,inf pour y en avoir
36. lt89 lt90 X/inf	prep,di,inf pour me conseiller
37. lt89 lt91 X/inf	prep,di,inf pour lui conseiller
38. lt89 lt85 X/inf	prep,di,inf pour y souper

39. lt89 lt86 X/inf	prep,di,inf pour en devoir
40. neg lt83 lt92 inf	neg,adv,di,inf ne pas le voir
41. lt92 inf	di,inf le voir
42. lt93 adjinf	prep,inf avant de tendre
43. lt94 nn	det,nn le pain
44. lt95 adj	det,adj le bel
45. lt95 adjnn	det,adjnn le creux
46. lt94 adj nn	det,adj,nn son bel homme
47. lt94 adjnn nn	det,adj,nn ton curieux oncle
48. det X nn	det,adj,nn l'actuel engagement
49. lt84 lt78 X/v	di,di,v me le lit
50. lt84 lt85 X/v	di,di,v m'y fabrique
51. lt84 lt86 X/v	di,di,v m'en bouche
52. lt78 lt87 X/v	di,di,v le lui lit
53. lt78 lt85 X/v	di,di,v les y plante
54. lt78 lt86 X/v	di,di,v l'en para
55. lt87 lt86 X/v	di,di,v leur en beurre
56. lt85 lt86 X/v	di,di,v y en aura
57. lt90 X/v	di,v me pique
58. neg lt84 X/v	neg,di,v ne vous critique
59. neg lt78 X/v	neg,di,v ne le ménage
60. neg lt87 X/v	neg,di,v ne leur joue
61. neg lt85 X/v	neg,di,v n'y entre
62. neg lt86 X/v	neg,di,v n'en finance
63. neg X	neg,v n'opéra
64. spr lt88 X/v	spr,di,v je le beurre

65. spr X	spr,v je voyage
66. lt96 neg X/v	spr,neg,v elle ne dément
67. lt96 di X/v	spr,di,v elle me distrait
68. lt97 lt98	prep,di sans moi, À elle
69. lt96 v	spr,v elle joue
70. det X/nn/adjnn lt88	det,nn,di mon mousseux me
71. lt95 X/nn/adjnn lt88	det,nn,di leur mousseux me
72. lt94 X/nn/adjnn neg	det,nn,neg certains mousseux ne
73. det X neg	det,nn,neg mes vaches ne
74. lt95 X/adjnn adj v	det,nn,adj,v leur mousseux formidable coûte
75. lt95 adj X/adjnn v	det,adj,nn,v leur nouveau mousseux a
76. lt94 X/adjnn adj adj v	det,nn,adj,adj,v ce mendiant rouennais minable dit
77. lt95 adj adj X/adjnn v	det,adj,adj,nn,v ce fantastique nouveau mousseux a
78. lt94 adj X/adjnn adj v	det,adj,nn,adj,v son nouveau lépreux rouennais dit
79. det X X v	det,adjnn,adjnn,v chaque soldat blessé attend
80. lt94 X X v	det,adjnn,adjnn,v les soldats blessés ont
81. det X X lt100	det,adjnn,adjnn,di chaque soldat blessé me
82. lt94 X X lt100	det,adjnn,adjnn,di le soldat blessé me
83. det X X neg	det,adjnn,adjnn,neg chaque soldat blessé ne
84. lt94 X X neg	det,adjnn,adjnn,neg le soldat blessé ne
85. det X X X v	det,adjnn,adjnn,adjnn,v chaque soldat grec blessé a
86. lt94 X X X v	det,adjnn,adjnn,adjnn,v les soldats grecs blessés ont
87. det X X X lt100	det,adjnn,adjnn,adjnn,di chaque soldat grec blessé me
88. lt94 X X X lt100	det,adjnn,adjnn,adjnn,di les soldats grecs blessés me
89. det X X X neg	det,adjnn,adjnn,adjnn,neg chaque soldat grec blessé ne
90. lt94 X X X neg	det,adjnn,adjnn,adjnn,neg les soldats grecs blessés ne

91. det X X X X v	det,adjnn,adjnn,adjnn,adjnn,v chaque simple soldat grec blessé a
92. lt94 X X X X v	det,adjnn,adjnn,adjnn,adjnn,v les simples soldats grecs blessés ont
93. det X X X X lt100	det,adjnn,adjnn,adjnn,adjnn,di chaque simple soldat grec blessé me
94. lt94 X X X X lt100	det,adjnn,adjnn,adjnn,adjnn,di les simples soldats grecs blessés me
95. det X X X X neg	det,adjnn,adjnn,adjnn,adjnn,neg chaque simple soldat grec blessé ne
96. lt94 X X X X neg	det,adjnn,adjnn,adjnn,adjnn,neg les simples soldats grecs blessés ne
97. lt95 X/adjnn adj lt97	det,nn,adj,prep leur mousseux formidable avec
98. lt95 adj X/adjnn lt97	det,adj,nn,prep leur nouveau mousseux pour
99. lt94 X/adjnn adj adj lt97	det,nn,adj,adj,prep ce mendiant rouennais minable sans
100. lt95 adj adj X/adjnn lt97	det,adj,adj,nn,prep ce fantastique nouveau mousseux par
101. lt94 adj X/adjnn adj lt97	det,adj,nn,adj,prep son nouveau lépreux rouennais parmi
102. det X X lt97	det,adjnn,adjnn,prep chaque soldat blessé par
103. lt94 X X lt97	det,adjnn,adjnn,prep les soldats blessés par
104. det X X X lt97	det,adjnn,adjnn,adjnn,prep chaque soldat grec blessé par
105. lt94 X X X lt97	det,adjnn,adjnn,adjnn,prep les soldats grecs blessés par
106. det X X X X lt97	det,adjnn,adjnn,adjnn,adjnn,prep chaque simple soldat grec blessé par
107. lt94 X X X X lt97	det,adjnn,adjnn,adjnn,adjnn,prep les simples soldats grecs blessés par
108. lt94 X/nn/adjnn (end/string)	det,nn l'avocat
109. lt97 det X/nn/adjnn (end/s)	prep,det,nn chez mon avocat
110. lt97 lt94 X/nn/adjnn (end/s)	prep,det,nn chez leur avocat
111. lt94 adj X/nn/adjnn (end/s)	det,adj,nn ce nouvel avocat
112. X neg v	adjnn,neg,v avocat ne fait
113. X v	adjnn,v avocat fait

3. At this point, any word which has been temporarily marked "ambiguous number" (bit 57 set to 1) is passed through a set to rules to determine whether it is actually singular or plural. This information is needed for the algorithm which adds the consonants of liaison (e.g., voix intérieure with no liaison; voix intérieures with /z/ of liaison).

4. In French as in English (e.g., Time flies like an arrow), there are sentences which cannot be analyzed unless the context is supplied. Since we do not analyze the context in a sentence such as the following, our rules can complete the analysis of only one word:

La	belle	ferme	le	voile.
(determiner)	(adjective/noun)	(verb/noun)	(determiner/pronoun)	(verb/noun)

The fact that we cannot complete the linguistic analysis without context does not interfere with the phonemic transcription (all the words are transcribed correctly), but on the prosodic level we do not know where to place the highest pitch and a slight pause:

La	belle			ferme	le	voile,
(determiner)	(noun)			(verb)	(determiner)	(noun)

La	belle	ferme			le	voile,
(determiner)	(adjective)	(noun)			(pronoun)	(verb)

B. RULES TO DISAMBIGUATE AMBIGUOUS SPELLINGS

General comments:

In English, there are cases in which the pronunciation of a word depends on its grammatical function and/or its semantic content: lives, read, lead, wind, tear, row, bow, sow, present, object, etc. Similar cases occur in French and we must conduct some kind of analysis to determine which pronunciation is to be used.

When the words belong to different grammar categories (président, inventions, fier, est, vis, etc.), a linguistic analysis is generally sufficient and can be 100% accurate; for example, in the case of président, we can compare the words which precede and follow président with the words in verb lists (ils, elles, la, les, ne, -ils, -elles, pas, etc.) and with the words in noun lists (un, ce, mon, ton, etc.) and easily arrive at a decision.

When the words belong to the same grammar category (e.g., fils, lacs), we can first try a linguistic analysis. For example, if one or several words from singular lists (mon, ce, aîné, etc.) are found before or after fils, we can conclude that fils is singular and must be pronounced /fis/. If the linguistic analysis does not succeed, we need a semantic analysis; by comparing the semantic environment of fils with the roots contained in two lists (one containing the roots frequently associated with the meaning threads and another containing the roots frequently associated with the meanings son/sons), we can count the number of successful matches in each list and determine the probable meaning of fils. If the lists are long and if we search several lines before the occurrence of fils, the probability of success can be very close to 100%. We do not search the text which follows the input string containing the ambiguous spelling since, normally, a writer provides the human readers with sufficient information by the time they reach the end of the breath-group which

contains the ambiguous spelling. Obviously, there are times when the writer wants to retain the ambiguity (e.g., to make a pun); for that reason, we give both pronunciations when our program does not solve the ambiguity except for the cases in which one pronunciation is far more frequent than the other (e.g., auto, est, forte, gens, papas, soit, y).

(For convenience, we use the term "ambiguous spelling" to designate any spelling which has more than one pronunciation. In that context, fils is an ambiguous spelling, but cours is not since, although it has several meanings, it has only one pronunciation.)

Our list of ambiguous spellings (Group Five in the lexicon) contains 158 items. Establishing that list has required a great amount of time and we believe it is practically complete; missed items, if any, are probably of the type "proper name/common name" (e.g., Lot, Damas, Eu, But) or of the type "slang/standard French" (e.g., flous).

We have not included words like frimas, ramas, etc. which some people pronounce with final /a/ (for the verb) or /ɑ/ (for the noun).

We have not included some of the words for which a single pronunciation is acceptable: but, lacèrent, lampas, mil, sas, sus.

a. But: Some people pronounce the common noun but with a final /t/, especially in sports: trois buts à deux (provided the liaison is not made). In our program, the common noun but is pronounced /by/ except in the expression de but en blanc which we have placed on the idiom list.

b. Lacèrent: Some persons pronounce this word with /ɑ/ when it is the past tense of lacer, with /a/ when it is the present tense of lacérer. Our program uses only /a/.

c. Lampas: Some persons pronounce this word with /s/ in cases other than the past tense of lamper. Our program uses only the form without /s/.

d. Mil: Some persons use /j/ instead of /l/ when this word is synonymous with millet. Our program uses only /mil/.

e. Sas: This word is sometimes pronounced without a final /s/ when it is synonymous with crible. Our program uses only /sas/ (this word is in the lexicon).

f. Sus: Some persons pronounce /sy/ for the verb and /sys/ for all other cases. We are using the pronunciation /sys/ only in a few expressions which we have placed on the idiom list.

If used without periods or with lower-case letters, many abbreviations would become ambiguous: ces (collège d'enseignement secondaire), cet (collège d'enseignement technique), coder (commission de développement économique régional), etc., but we have not found any such cases.

Usually, an ambiguous spelling (not already ending with s) is no longer ambiguous when the letter s is added; for example, reporter, lut,

parent are ambiguous, but reporters, luts, parents are not. Those non-ambiguous spellings must be caught by our program before the final s is removed; some are caught by the lexicon (e.g., luts, posters, reporters), and the others are caught by the table of endings (e.g., all those ending with ents: évidents, ferments, présidents, etc.). The spellings which are ambiguous with a final s and without a final s (e.g., auto/s, bêta/s) need to be listed only in the singular form.

As explained on pages 28-29, words which contain one of the prefixes on our list plus an ambiguous spelling (e.g., désinfections) also go through the disambiguation process (except for the words listed on page 29).

Frequency of ambiguous spellings:

The following texts were chosen at random:

1. L'Etranger (Albert Camus): 2,000 initial words of Chapter One
2. Le Canard Enchaîné (March 21, 1984): Article entitled: "C'est la fête à Gaston"
3. La Voix du Combattant (April 1984): Article entitled: "Drogues"
4. Le guide 1979 de vos droits et démarches: Section entitled: "L'apprentissage"

In the first text, we find thirty-two ambiguous spellings out of 2,000 words. They are est (twenty-four times), plus (five times), tous (once), content (once), and gens (once).

In the second text, we find nine ambiguous spellings out of 532 words. They are plus (five times) and est (four times).

In the third text, we find sixteen ambiguous spellings out of 977 words. They are est (eight times), plus (seven times), and tous (once).

In the fourth text, we find twelve ambiguous spellings out of 773 words. They are est (eleven times) and plus (once).

Two facts stand out from our analysis of four randomly chosen texts:

1. Although the texts are rather short, the proportions of ambiguous spellings are remarkably similar: 1.6%, 1.7%, 1.6%, and 1.6%.

2. In each text, the most frequent ambiguous spellings are est and plus. Out of a total of 69 ambiguous spellings, est is used 47 times and plus is used 18 times. Since the most frequent uses of est (il est, elle est, c'est, n'est, etc.) and some of the combinations with plus (e.g., de plus en plus) are on the idiom list, 25 of the 47 uses of est and 5 of the 18 uses of plus are found in the lexicon. Even if we had no disambiguation rules at all and transcribed all remaining est and plus with /ɛ/ and /ply/, only one item (tous in the first text) would be transcribed incorrectly; the error rate, at the word level, would be 1/4,282 or about two words in ten thousand. This error rate may seem insignificant, but —as already stated— the purpose of this program is to try to equal the performance of an educated reader and we have disambiguation rules for

each of the ambiguous spellings on our list.

Rules:

Our program must attempt to analyze correctly the input strings which have two or three consecutive ambiguous spellings as well as those which have individual ambiguous spellings:

Les amis de ce président admettent que ...
Les amis de ce président excellent disent que ...
Les amis de ce président excellent et disent que ...
Les amis de ce président excellent et content président ...
Les amis de ce président excellent content une histoire ...

[Consecutive ambiguous spellings occur rarely, but —since human readers experience no difficulty— our program ought to interpret them correctly.]

Printing all our disambiguation rules would require nearly one hundred pages. Therefore, we will discuss only three typical rules.

1. Category: couvent

This rule deals with the thirteen ambiguous ent spellings (page 138) which can be verbs or nouns. N means that the word is to be given the pronunciation and the features listed in first position on page 138; V means that the word is to be given the pronunciation and the features listed in second position.

V if followed by a hyphen and a word from list 104
V if followed by a word from list 83
V if preceded by a word from lists 84, 101, 104, or a word marked (neg)
V if preceded by a word marked (plural)
V if preceded by a word marked (number)
V if preceded by a word marked (feminine)
V if preceded by a word marked (neg) and a word from list 105
V if preceded by a word from list 84 and a word from list 105
V if preceded by a word from list 104 and a word from list 105
V if preceded by a word from list 109
V if preceded in first position by a noun
V if only one word in input string
V if preceded in second, third, fourth or fifth position by a word from list 107 (e.g., trente et un président)
N if preceded by a word from lists 102, 103
N if preceded by a word from list 102 and a word from list 95
N if preceded by a word from list 110 and a word from list 95
N if preceded by a word marked (determiner,singular)
N if followed by a word marked (adjective)
N if followed by a word marked (adjective/noun)
N if followed by a word marked (verb)
N if preceded by a word marked (verb)
N if preceded by a word marked (adverb)
N if preceded by a word from list 103 and a word marked (adjective) or marked (adjective/noun)

N if preceded by a word from list 103, a word marked (adjective) or marked (adjective/noun), and a word marked (adjective) or marked (adjective/noun)
N if preceded by a hyphen

2. Category: convient

This rule deals with convient, obvient, pressent. S means that the word is to be marked (verb,singular) and that the first pronunciation shown on page 138 is to be used; P means that the word is to be marked (verb,plural) and that the second pronunciation is to be used.

[Only three verbs are in this category, but this is a case in which 100% accuracy is difficult to achieve; in particular, our rules do not analyze correctly a sentence such as "le professeur de ces étudiants et de ces étudiantes pressent les difficultés ...".]

P if followed by a hyphen and a word from list 104
P if preceded by a word from list 104
P if preceded by a word marked (number)
P if preceded by a word marked (plural)
P if preceded by a word from list 104 and a word marked (neg)
P if preceded by a word from list 104 and a word from list 88
P if preceded by a word from list 104 and a word marked (neg) and a word from list 88
P if preceded by a word from list 104 and a word from list 88 and a word from list 88
P if preceded by a word from list 104 and a word marked (neg) and a word from list 88 and a word from list 88
S if followed by a hyphen and a word from list 106
S if preceded by a word from list 106
S if preceded by a word from list 106 and a word marked (neg)
S if preceded by a word from list 106 and a word from list 88
S if preceded by a word from list 106 and a word marked (neg) and a word from list 88
S if preceded by a word from list 106 and a word from list 88 and a word from list 88
S if preceded by a word from list 106 and a word marked (neg) and a word from list 88 and a word from list 88
S if followed by a word marked as subordinate
P if preceded in second, third, fourth or fifth position by a word from list 107
P if preceded by a plural word
P if preceded by a plural word and a word marked (neg)
P if preceded by a plural word and a word from list 88
P if preceded by a plural word and a word marked (neg) and a word from list 88
P if preceded by a plural word and a word from list 88 and a word from list 88
P if preceded by a plural word and a word marked (neg) and a word from list 88 and a word from list 88
S if preceded by a singular word (or proper name)
S if preceded by a singular word (or proper name) and a word

marked (neg)
S if preceded by a singular word (or proper name) and a word from list 88
S if preceded by a singular word (or proper name) and a word marked (neg) and a word from list 88
S if preceded by a singular word (or proper name) and a word from list 88 and a word from list 88
S if preceded by a singular word (or proper name) and a word marked (neg) and a word from list 88 and a word from list 88

3. Category: fils

Some of the problems posed by fils are solved before we reach this point in our program; for example, beaux-fils is rewritten in Step 5, and compte-fils, petit-fils, père et fils, etc. are on the idiom list.

The semantic analysis is to be done only if the linguistic analysis fails to resolve the ambiguity.

Linguistic analysis:

The word fils is marked (noun,singular) and transcribed /fis/

- if it is preceded in first, second, or third position by a word from list 103 or list 108
- or if it is followed or preceded in first or second position by a singular adjective or a singular adjective/noun
- or if it is immediately followed or preceded by a singular noun

Note that a singular verb does not necessarily mean that fils is singular (e.g., un assortiment de fils est vendu ...).

Semantic analysis:

Our program searches the input string and ten preceding breath-groups (or as many as are available if ten are not available) for words beginning with the roots listed in the following two sets; the roots which are marked with an asterisk are counted only if they constitute the whole word.

When the search is completed, fils is marked (nn,m,pl) and is transcribed /fil/ if more roots are found from the first set than from the second set; if the reverse is true, fils is marked (nn,m,an) and is transcribed /fis/. If the numbers are equal or if no roots are found, fils is marked (nn,m) and both transcriptions are given[15].

First set:

aerj	agraf	ãfil	alyminj	ãmɛl	arɛɲ	ariko
asje	ataʃ	barbøl	bati	bɛʒ	blã	blø
blɛ̃d	bɔbin	brãʃ	brɛ̃	brɔ̃	ʃãvr	ʃif
de*	demɛl	devid	eʃøvo	egɥij	ekry	elɛktr
epɛ̃gl	etir	fɛr	fɔ̃se	fyzo	galvan	gri

ʒon	ʒyt	kɑbl	kard	kənuj	klɛr	kɔrd
kɔt	ku*	kud	kulœr	kup	kut	kuzɑ̃
kuze	kuzɛ	kuzjɔ̃	kuzy	kɥivr	lɛn	lɛtɔ̃
lɛ̃	lɛ̃ʒ	ligat	liɲø	marjɔn	marɔ̃	mɛrs
metal	metje	mov	nwar	ɔr*	ɔrizɔ̃	pɛɲ
pɛ̃ist	pik	pwɛ̃	rakɔr	rapj	rəpriz	roz
ruʒ	rwɛ	sizo	supl	swa	syspɑ̃	taj
tɑ̃d	tɑ̃sj	tɛ̃t	telef	teleg	tɔrd	tɔrsa
trefil	urd	url	vɛr*	vɛrtik	vjɛrʒ	vjɔlɛ

Second set:

adɔpt	adylt	ɑ̃fɑ̃	ajø	ajœl	aku∫
aljɑ̃s	asɑ̃d	ɑ̃sɛ	batar	batɛm	batiz
bebe	bɛ̃ʒamɛ̃	bɛrs	bibørɔ̃	desɑ̃d	dofɛ̃
ekɔl	ene	epu	erit	fam	fij
frɛr	fwaje	go*	gars	gɑte	gɔs
guvɛrn	isy	ʒɑ̃dr	ʒwɛ	ʒymo	kadɛ
ku∫	kuzɛ̃	kuzin	leʒiti	liɲe	maʒ
mamɑ̃*	marɛ	mari	marj	mɛr*	minœ
natyrɛl	ne*	nɛsɑ̃	nøvø	nɔ̃*	nuris
œvrø	ɔrfølɛ̃	papa	parɑ̃	parɛ̃	patri
pɛr*	pɔsty	prenɔ̃	puine	rəkɔn	sɑ̃*
saʒ	sœr	su∫	ynik		

The combination sans fils is transcribed with /fil/ when the roots for haricots, télégraphie, or téléphone are present; it is transcribed /fis/ in other cases. The combination fils de la Vierge is transcribed /fis/ when fils is found to be singular and is transcribed /fil/ in other cases.

C. FINAL STEP IN THE ASSIGNMENT OF LINGUISTIC FEATURES

If some ambiguity had to be resolved and if the process was successful, we repeat the first step since it is possible to assign additional linguistic features after ambiguities have been resolved.

D. SEPARATION OF INPUT STRING INTO BREATH-GROUPS

General comments:

1. In French, the number of syllables is the same as the number of vowel phonemes; therefore, our algorithm counts the number of vowels.

2. If the number of vowels is greater than 14, the program looks at the whole input string to determine whether one of the rules given below can be applied. If this search is successful and if, after the rule is applied, there are still one or several sequences over 14 syllables long, the program starts at rule 1 again, and so on until no sequences with over 14 syllables are left or until no rules can be applied.

3. The rules must be applied in the order given below. However:

a. a rule is not applied if it would result in a breath-group having fewer than three syllables; the program proceeds to the next rule.

b. a rule is skipped if the ratio of the number of syllables in two adjoining groups is greater than 2. The program looks for a rule which gives a better ratio (except that rules 31, 32, and 33 cannot be substituted for rule 30 and that rules 37, 38, and 39 cannot be substituted for rule 36); if it does not find one, it returns to the rule which had been skipped. For example, rule 13 would divide the following string after déranger and would create a string with 12 syllables and another with only 4 syllables (ratio 12/4=3); we therefore skip rule 13 and find that rule 38 divides the string after efforceront and that the ratio falls to 1.66:

Elles s'efforceront de ne pas déranger la réunion.

Rules:

Separation into breath-groups occurs under the following conditions:

1. Between word marked (noun) or (proper name) and word marked (neg)

   les deux nouveaux films // n'ont pas encore été présentés

2. Between word marked (adjective/noun) and word marked (neg)

   deux des soldats grecs // n'ont pas assisté à la réunion

3. Between word marked (adjective) and word marked (neg)

   tous les films récents // ne me paraissent pas très intéressants

4. Between word marked (noun) or (proper name) and word from list 111

   les deux nouveaux romans // auxquels vous pensez ne valent rien

5. Between word marked (adjective/noun) and word from list 111

   les deux jeunes soldats grecs // auxquels vous pensez sont là-bas

6. Between word marked (adjective) and word from list 111

   les enfants intelligents // auxquels vous pensez sont partis

7. Before combination of word from list 102 + word from list 104

   nous avons travaillé // avec quelques-unes de vos amies

8. Before combination of word from list 102 + word from list 106

   nous avons travaillé // avec celle-ci sans difficulté

9. Before combination of word from list 102 + word from list 109

nous ne savons plus // par où nous sommes passés pour venir ici

10. Before combination of word from list 102 + word from list 110

nous avons travaillé // pendant qu'ils s'amusaient à Paris

11. Before word/idiom marked (subordinate)

nous partirons tous // à moins qu'il ne commence à pleuvoir

12. Between word marked (infinitive) and word from list 102

nous devrons nous séparer // dès l'arrivée de mes parents

13. Between word marked (infinitive) and word marked (determiner)

nous devrons accepter // les inconvénients de la situation
nous devrons nous habituer // aux inconvénients de la situation

14. Between word marked (past participle) and word from list 102

nous avons tous attendu // sans grand espoir de les revoir

15. Between word marked (past participle) and word marked (determiner)

nous avons encore perdu // l'adresse de la société
nous avons tous perdu // des sommes considérables ici

16. Between word marked (verb) and word from list 102

nous nous séparerons // dès l'arrivée de votre famille

17. Between word marked (verb) and word marked (determiner)

nous accepterons // les inconvénients de la situation
nous nous habituerons // aux inconvénients de la situation

18. Before combination of word from list 86 and word marked (present participle)

je me suis trompé de chemin // en essayant d'arriver très tôt

19. Between word marked (noun) or (proper name) and word marked (verb)

les deux nouveaux romans // seront publiés demain matin

20. Between word marked (adjective/noun) and word marked (verb)

les jeunes soldats grecs // ont assisté à la réunion

21. Between word marked (adjective) and word marked (verb)

tous les films récents // paraissent complètement imbéciles

22. Between word marked (noun) or (proper name) and word from list 88

les deux nouveaux romans // vous seront envoyés demain soir

23. Between word marked (adjective/noun) and word from list 88

les deux soldats grecs // y ont assisté jusqu'à midi et quart

24. Between word marked (adjective) and word from list 88

tous les films récents // me paraissent complètement idiots

25. Between word marked (number) and word marked (verb)

les alliés en 1920 // évacuèrent la Rhénanie

26. Between word marked (number) and word from list 88

les alliés en 1920 // se retirèrent de cette zone

27. Between word marked (verb) and word marked (verb)

celle que vous désirez // coûte certainement trop cher

28. Between word marked (singular) and word marked (plural)

les membres du parti démocrate // réunis à Philadelphie

29. Between word marked (plural) and word marked (singular)

le syndicat des ouvriers // fort de ses droits inaliénables

30. After a sequence of four words marked (adverb) except if they occur before an infinitive, a past participle, or an adjective

il n'écoute pas toujours très sérieusement // les discours du chef

31. After a sequence of three words marked (adverb) except if they occur before an infinitive, a past participle, or an adjective

il n'écoute pas très sérieusement // les discours du nouveau chef

32. After a sequence of two words marked (adverb) except if they occur before an infinitive, a past participle, or an adjective

il écoute très sérieusement // les discours du nouveau chef

33. After a word marked (adverb) except if it occurs before an infini-

tive, a past participle, or an adjective

il écoute sérieusement // les discours du nouveau chef

34. Before a word from list 89

il travaillera plus dur // afin de réussir la prochaine fois

35. Before a word from list 97

il réussira certainement // grâce à votre aide généreuse

36. Before a word marked (coord)

nous aurions visité Philadelphie // et les villes voisines

37. Before a sequence of two words marked (preposition)

il n'aurait pas encore réussi // d'après votre opinion

38. Before a word marked (preposition)

nous aurions visité Philadelphie // avant les villes voisines

39. Before a word marked (determiner)

nous visiterons Philadelphie // la semaine prochaine

40. After a word marked (infinitive)

lui et moi préférons travailler // individuellement

41. After a word marked (past participle)

nous avons généralement travaillé // individuellement

42. After a word marked (verb)

lui et moi travaillerons // souvent individuellement

43. No division if none of the above rules can be applied

certaines nouvelles dispositions anticonstitutionnelles

E. PRONUNCIATION OF NUMBERS ENDING WITH 5, 6, 8, 10

There is much hesitation in France about the pronunciation of these numbers; the rules given below are generally acceptable among educated speakers.

The rules apply to all numbers ending with /sɛ̃k/, /sis/, /ųit/, and /dis/, that is to words which have been assigned the feature "tn" (bit

70 set to 1).

1. Cinq and huit:

/sɛ̃k/ and /ɥit/ drop their final consonant when they are followed by a noun, adjective, or adjective/noun beginning with a consonant.

2. Six and dix:

/sis/ and /dis/ drop their final consonant when they are followed by a noun, adjective, or adjective/noun beginning with a consonant. They change their final /s/ to /z/ when they are followed by a noun, adjective, or adjective/noun beginning with a vowel or semivowel.

F. ADDING THE CONSONANTS OF LIAISON

General comments:

1. The word grand, by itself, is pronounced /grɑ̃/. The string un grand ami is pronounced /œ̃grɑ̃tami/; the additional /t/ is referred to as a consonant of liaison. Today[16], the only consonants which can be consonants of liaison are /t/, /z/, /n/, /r/, /p/ and /g/. Examples:

/t/	petit homme	/pətitɔm/
	grand homme	/grɑ̃tɔm/
/z/	les idées	/lezide/
	deux amis	/døzami/
	chez elle	/ʃezɛl/
/n/	on attend	/ɔ̃natɑ̃/
/r/	premier étage	/prəmjeretaʒ/
/p/	trop âgé	/tropaʒe/
/g/	long hiver	/lɔ̃givɛr/

The semivowel /j/ is used as a sound of liaison in the compound nouns gentilhomme and gentilhommière.

Liaison should not be confused with linking. Linking occurs when a word ending with a consonant sound is followed by a word beginning with a vowel or semivowel sound: thus, the spoken syllables of grande amie are /grɑ̃ dɑ mi/, not /grɑ̃d a mi/.

2. There is no liaison in neuf ans, neuf heures, dix-huit, etc.: in those examples, we simply have a voiceless consonant (/f/, /s/) which becomes voiced (/v/, /z/). All such expressions are on our list of idioms.

3. Native speakers of French do not agree about which words denasalize their final nasal vowels in the liaison process. We have adopted the pronunciations which are the most frequent among educated speakers in northern France. Examples: /œ̃bɔnami/, but /mɔ̃nami/.

4. There is also disagreement about the pronunciation of words ending in er when the liaison is made (e.g., premier étage): some speakers keep

/e/ while others tend to pronounce a sound which is close to /ɛ/. In this program, we use /e/: /pɒrɔmjeretaʒ/.

5. Many types of liaisons have been placed on the list of idioms (e.g., the clusters -t-il, -t-elle, -t-on, expressions such as de temps en temps, accent aigu, tout à coup, fait accompli, cas échéant, corps et âme).

6. If we reach this part of our program with the sequence singular adjective/noun with potential liaison + singular adjective/noun beginning with a vowel or semivowel (e.g., un malheureux envieux), we do not have sufficient information to decide whether the liaison should be made or not. Since, in this case, it is preferable not to make a liaison rather than make a wrong liaison and since, in any case, the sequence adjective/noun + adjective/noun more often represents noun + adjective than adjective + noun, our program does not add a consonant of liaison.

If both adjectives/nouns are plural (e.g., les malheureux envieux), we make the liaison since it is either compulsory (adjective + noun) or optional (plural noun + plural adjective).

7. Although errors are extremely rare and do not affect our accuracy level in any significant way, it is still regrettable that we cannot perform the type of contextual analysis that would prevent us from making the liaison in the first sentence of these three pairs:

Vous travaillez trop ici. (No liaison since trop modifies the verb)
Vous arrivez trop en retard. (Liaison since trop modifies en retard)

On est toujours le sujet du verbe. (No liaison)
On est toujours le sujet de quelqu'un. (Liaison)

Nous sommes bien en France. (No liaison if bien modifies sommes)
Nous sommes bien en France. (Liaison if bien modifies en France)

8. When we reach this point in our program, the necessary data for the liaison has been stored in two ways:

a. All the determiners, numbers, pronouns, prepositions, adverbs, conjunctions, and some verbal forms with which we believe a liaison should be made are in the lexicon. If one of those words occurs in an input string, the bit corresponding to the potential liaison is set: bit 65 for /z/, bit 66 for /n/, bit 67 for /t/, and bit 68 for /p/.

b. The features "sxzend"", "dtend", "nend", "rend", and "gend" are attached to the words ending with "s, x, or z", "d or t", "n", "r", "g" when they enter the transcriber (page 19).

We need this double system because not all prepositions, adverbs, conjunctions, etc. ending with d, n, p, s, t, x, or z have a compulsory liaison: for example, we make the liaison with jamais, but not with longtemps even though both are adverbs ending with s. Another reason is that it would not be practical to list in the lexicon all the forms of all the adjectives, adjectives/nouns, nouns, etc. with which a liaison can be made.

9. In all the rules below, we assume that the word being considered is followed by a word beginning with a vowel or semivowel sound. No liaison is made with "hiatus words" (e.g., héros, hache) or across guillemets (words for which bit 76 is set to 1).

10. We attach the consonant of liaison to the second word; for example, the sequence /le ide/ becomes /le zide/, not /lez ide/.

Rules:

Rules 1 through 20 concern words which are not preceded or followed by a hyphen (bits 72 and 73 set to 0).

Rule 1: Add the liaison if —for the word being considered— bit 6 is equal to 1 and the sum of bits 65 through 68 is equal to 1, and if —for the next word— the sum of bits 1, 2, 3, 7 through 17, and 20 through 51 is equal to 0.

This means, for example, that the determiner leurs will add a /z/ of liaison if NOT followed by a finite verb, an imperative, an infinitive, a preposition, a present participle, etc. Another way to express the rule is to say that leurs adds the consonant of liaison if followed by a noun, an adjective, an adjective/noun, a proper name, OR a word which has not been assigned any grammar category (bits 1 through 51 set to 0). [This rule also includes the liaison in cases such as les USA, aux USA.]

Rule 2: Add the liaison if —for the word being considered— bit 10 is equal to 1, bit 73 is equal to 0, and the sum of bits 65 through 68 is equal to 1, and if —for the next word— bit 1 is equal to 1, and if —bit 16 of the preceding word being equal to 1— the sum of bits 10, 11, and 12 for the second preceding word is equal to 0.

This means, for example, that the subject pronoun on will add a /n/ of liaison if not preceded by a hyphen and if followed by a verb. It also means that a liaison will take place in Elle travaille et vous êtes là, but will not take place in Lui et vous êtes fatigués.

Rule 3: Add the liaison if —for the word being considered— bit 10 is equal to 1, bit 73 is equal to 0, and the sum of bits 65 through 68 is equal to 1, and if —for the next word— bit 11 is equal to 1, and if —bit 16 of the preceding word being equal to 1— the sum of bits 10, 11, and 12 for the second preceding word is equal to 0.

This means, for example, that the subject pronoun on will add a /n/ of liaison if not preceded by a hyphen and if followed by a pre-verb object pronoun (e.g., en). It also means that a liaison will occur in Il en parle et vous en parlez, but will not take place in Lui et vous en parlez.

Rule 4: Add the liaison if —for the word being considered— bit 11 is equal to 1, bit 73 is equal to 0, and the sum of bits 65 through 68 is equal to 1, and if —for the next word— bit 11 is equal to 1.

This means, for example, that the object pronoun les will add a /z/ of

liaison if not preceded by a hyphen and if followed by a pre-verb object pronoun (e.g., y).

Rule 5: Add the liaison if —for the word being considered— bit 11 is equal to 1, bit 73 is equal to 0, and the sum of bits 65 through 68 is equal to 1, and if —for the next word— bit 1, or bit 2, or bit 3 is equal to 1.

This means, for example, that the object pronoun les will add a /z/ of liaison if not preceded by a hyphen and if followed by a verb (e.g., il les entend), by an imperative (e.g., ne les écoute pas), or by an infinitive (e.g., je vais les écouter).

Rule 6: Add the liaison if —for the word being considered— bit 7 is equal to 1 and the sum of bits 65 through 68 is equal to 1, and if —for the next word— the sum of bits 1, 2, 7 through 10, 13 through 16, 20 through 51 is equal to 0.

This means, for example, that the preposition sans will add a /z/ of liaison if NOT followed by a finite verb, an imperative, an adverb, a past participle, etc. Another way to express the rule is to say that sans adds the consonant of liaison if followed by an infinitive, a noun, an adjective, a determiner, an object pronoun, a present participle, a proper name, an adjective/noun, OR by a word which has not been assigned any grammar category (bits 1 through 51 set to 0).

Rule 7: Add the liaison if —for the word being considered— bit 9 is equal to 1 and the sum of bits 65 through 68 is equal to 1, and if —for the next word— the sum of bits 1, 2, 4, 8, 10, 13, 14, 16, 17, 20 through 51 is equal to 0.

This means, for example, that the adverb pas will add a /z/ of liaison if NOT followed by a finite verb, an imperative, a noun, a present participle, etc.

[We do not make the liaison with pis (except in the idiom pis-aller) and tôt (except in the idiom tôt ou tard).]

Rule 8: Add the liaison if —for the word being considered— bit 14 is equal to 1 and the sum of bits 65 through 68 is equal to 1, and if —for the next word— the sum of bits 1, 2, 3, 8, 12, 13, 14, 16, 20 through 51 is equal to 0.

This means, for example, that the conjunction quand will add a /t/ of liaison if NOT followed by a finite verb, an imperative, an infinitive, a disjunctive pronoun, etc.

Rule 9: Add the liaison if —for the word being considered— bit 16 is equal to 1 and the sum of bits 65 through 68 is equal to 1, and if —for the next word— the sum of bits 8, 12, 13, 14, 16, 20 through 51 is 0.

This means, for example, that the conjunction mais will add a /z/ of liaison if NOT followed by a finite verb, an imperative, an infinitive,

a disjunctive pronoun, etc.

Rule 10: Add the liaison if —for the word being considered— bit 1 is equal to 1 and the sum of bits 65 through 68 is equal to 1, and if —for the next word— the sum of bits 1, 2, 3, 8, 10, 11, 13, 14, 17, 20 through 51 is equal to 0.

This means, for example, that the verb suis will add a /z/ of liaison if NOT followed by a finite verb, an imperative, an infinitive, etc.

Rule 11: Add the liaison if —for the word being considered— bit 13 is equal to 1 and the sum of bits 65 through 68 is equal to 1, and if —for the next word— the sum of bits 1, 2, 3, 7 through 17, 20 through 51 is 0.

This means, for example, that vingt adds a /t/ of liaison if NOT followed by a finite verb, an imperative, an infinitive, a preposition, etc.

Rule 12: Add the liaison if —for the word being considered— bit 24 is equal to 1, bit 73 is equal to 0, and the sum of bits 65 through 68 is equal to 1, and if —for the next word— the sum of bits 6, 8, 9, 10, 12, 13, 14, 15, 16, 17, 20 through 51 is equal to 0.

This means, for example, that —if we have been unable to determine the grammar category of les— the liaison will still be made if les is followed by a verb, a noun, a preposition, etc.

Rule 13: Add the liaison if —for the word being considered— bit 25 is equal to 1, bit 73 is equal to 0, and the sum of bits 65 through 68 is equal to 1, and if —for the next word— the sum of bits 7 through 11, 14 through 16, and 20 through 51 is equal to 0.

This means that —if we have been unable to determine the grammar category of en— the liaison will still be made if en is followed by a verb, an imperative, an infinitive, etc.

Rule 14: Add the liaison if —for the word being considered— bit 30 is equal to 1, bit 73 is equal to 0, and the sum of bits 65 through 68 is equal to 1, and if —for the next word— the sum of bits 4 through 10, 12 through 16, and 18 through 51 is equal to 0.

This means, for example, that —if we have been unable to determine the grammar category of nous— the liaison will still be made if nous is followed by a verb, an imperative, an infinitive, etc.

Rules 15-20 are for combinations of adjectives, adjectives/nouns, and nouns; they use the features "sxzend" (bit 59), "dtend" (bit 60), "nend" (bit 61), "rend" (bit 62), and "gend" (bit 63).

Rule 15: Add a /z/ of liaison if a plural adjective, or adjective/noun, or noun is followed by a plural adjective, adjective/noun, or noun:

autres anciens amis	amis anglais
leurs vieux amis	des nez énormes

les amoureux astucieux	vulgaires adversaires
grosses oranges (sequence /s+z/)	mauvaises années (sequence /z+z/)

Rule 16: The transcribed strings /kur/, /lur/, and /fɔr/ do not have a liaison if marked singular adjective or singular adjective/noun (bit 55 set to 1: bits 5 or 19 set to 1). This means, for example, that there will be no liaison in the sequence un court arrêt.

Rule 17: The transcribed strings:

ɑ̃sjɛ̃ bɔ̃ sɛrtɛ̃ divɛ̃ mwajɛ̃ plɛ̃ prɔʃɛ̃ vɛ̃ vilɛ̃

become ɑ̃sjɛn bɔn sɛrtɛn divin mwajɛn plɛn prɔʃɛn vɛn vilɛn

if marked singular adjective (bits 5 and 55 set to 1) and if followed by a noun or an adjective (bits 4 or 5 set to 1) beginning with a vowel or semivowel sound (e.g., en plein hiver, un certain autre agent).

Rule 18: Bit 60 is reset to 0 when it is assigned to a singular adjective ending with /t/; bit 62 is reset to 0 when it is assigned to a singular adjective ending with /r/. This is to avoid that sequences such as cher ami, net avantage be pronounced /ʃɛrrami/ and /nɛttavɑ̃taʒ/.

Rule 19: Liaison occurs in the sequence singular adjective + noun:

bit 59: add /z/ (e.g., gros homme, faux ami)
bit 60: add /t/ (e.g., grand homme, petit âne)
bit 61: add /n/ (e.g., commun effort, malin esprit)
bit 62: add /r/ (e.g., premier étage)
bit 63: add /g/ (e.g., long hiver)

Rule 20: Liaison occurs in the sequence singular adjective + singular adjective if the sequence is final of string or if it is followed by a noun, a verb form, a pronoun, an adverb, or a word marked (neg).

Rule 21: This rule is for words which have bit 72 ("prehyph") set to 1 (word 1) and which are followed by a word (word 2) beginning with a vowel or semivowel sound:

A /t/ of liaison is added if bit 60 ("dtend") of word 1 is set to 1:

pet-en-l'air	peut-être	pied-à-terre	acquit-à-caution
grand-oncle	prend-elle	part-il	partent-ils
vendent-elles	guet-apens	mort-aux-rats	tout-à-l'égout

A /z/ of liaison is added if bit 59 ("sxzend") of word 1 is set to 1 and if word 2 is /i/ or /ɛ̃/ or if word 1 ends with a vowel other than /ɛ/ while word 2 is other than /a/ or /o/. Examples:

porte-les-y	manges-en	sors-en	petits-enfants	Champs-Elysées
beaux-arts	sans-abri	bas-Empire	sous-entendu	lès-Elbeuf

The words /ɑ̃/, /bjɛ̃/, /bɔ̃/, /mɔ̃/, /plɛ̃/, /mwajɛ̃/ become /ɑ̃n/, /bjɛn/,

/bɔn/, /nɔn/, /plɛn/, /mwajɛn/. Examples:

en-avant bien-être non-activité plein-emploi moyen-âge

Exceptions (e.g., quatre-vingt-un, mot-outil, Basses-Alpes, guets-apens, sud-est) are on the idiom list. A discussion of this rule is on page 179.

G. TREATMENT OF UNSTABLE VOWEL

We must now decide which /ɸ/ should be changed to /ə/ and which should be dropped. The consonants (C) are /p/, /t/, /k/, /b/, /d/, /g/, /f/, /s/, /ʃ/, /v/, /z/, /ʒ/, /m/, /n/, /ɲ/, /ŋ/, /l/, /r/, and /ˊ/ (initial phoneme of hiatus words). The rules must be applied in the order given below.

The combinations de ce que and ce que are on the list of idioms.

Rule 1: Change /ɸ/ to /ə/ if it occurs in a monosyllabic word before a word in which bit 76 is set to 1: à ce <<nouveau>> décor /a sə nuvo dekɔr/.

Rule 2: Change /ɸ/ to /ə/ if it occurs in /lɸ/ after hyphen: /di lə bjɛ̃/; remove /ɸ/ if it occurs in /ʒɸ/ or /sɸ/ after hyphen: /par u syi ʒ pɑse/, /ɛ s lyi u ɛl/.

Rule 3: In the combinations /ʒɸ nɸ/ + C, /dɸ nɸ/ + C, /sɸ nɸ/ + C, change the first /ɸ/ to /ə/ and remove the second /ɸ/: /ʒə n vø pɑ/. Both /ɸ/ become /ə/ if /nɸ/ precedes /lj/, /nj/, or /rj/: /esɛ də nə rjɛ̃ dirɸ/.

Rule 4: In the combination /ʒɸ tɸ/ + C, remove the first /ɸ/ and change the second /ɸ/ to /ə/: /kɑ̃ ʒ tə di/.

Rule 5: Change /ɸ/ to /ə/ if it occurs in the initial syllable of the breath-group before consonant: /lə ʃɸmɛ̃/, /krəve sa/.

Rule 6: Change /ɸ/ to /ə/ if it occurs in non-final position in a word after two or more consonants: /travɛrsəra/.

Rule 7: Change /ɸ/ to /ə/ if it occurs before consonant in final position after /bl/, /br/, /dg/, /dn/, /dr/, /fl/, /fr/, /ft/, /gl/, /gm/, /gr/, /kl/, /km/, /kr/, /ks/, /kt/, /lb/, /lʃ/, /lf/, /lg/, /lk/, /lm/, /lp/, /ls/, /lt/, /lv/, /mn/, /pl/, /pr/, /ps/, /pt/, /sk/, /sm/, /sp/, /st/, /tʃ/, /tm/, /tr/, /vr/: /lə ʃɸmɛ̃ travɛrsəra lɸ gufrə dy djɑblɸ/.

Rule 8: Change /ɸ/ to /ə/ if it occurs in monosyllabic /kɸ/: /me kə sɸ pɑsɸ til/, /e kə di ty/. Thus, we keep /ə/ even when it introduces a subordinate: Je crois que tu as tort.

Rule 9: Change /ɸ/ to /ə/ if it occurs before /ˊ/: /se lə ˊɜɛɸ u/, /sɑ̃ sə ˊero/, /kɜ̃te də ˊqit a vɛ̃/, /katrə ˊjaurt/. However, in 38, 48, 58, 68, 1008, 1011 and the corresponding ordinal numbers, /q/ and /ɜ/ are not preceded by /ˊ/ and /ɸ/ is dropped; these numbers are on the list of idioms.

Rule 10: Change /ɸ/ to /ə/ if it occurs in /sɸ/, /dɸ/, /ʒɸ/, /lɸ/, /mɸ/,

/nɸ/, /kɸ/, /tɸ/ before vowel: /də ɑ a be/, /sə a kwa ty pɑ̃sɛ/, /də œ̃ a sɑ̃/, /ʃwazise lə œ̃/, /syr sə il e parti/.

Rule 11: Change /ɸ/ to /ə/ if it occurs in initial syllable of word after a capital letter: /məsjɸ ləfɔr va a sədɑ̃/.

Rule 12: Change /ɸ/ to /ə/ if it occurs in a monosyllabic word in final position of breath-group: /ytilize l artiklə lə/, /lize lɸ mo ʒə/.

Rule 13: Remove /ɸ/ in final position of breath-group.

Rule 14: Change /ɸ/ to /ə/ if it occurs in initial syllable of word after a word ending with /ə/: /uvrə sə livr/.

Rule 15: Change /ɸ/ to /ə/ if it occurs in final position of word before /lj/, /rj/, /nj/, /rw+V/, /lɥ+V/, /rɥ+V/: /də sə ljɸ/, /vu nɸ fɛtə rjɛ̃/, /il esɛ də nje/, /prane sə rwaʒ/, /parlɔ̃ də lɥi/, /sɥive sə rɥiso/.

Rule 16: Change /ɸ/ to /ə/ if it occurs in final position of word before a hyphenated word containing only one vowel sound (excluding /ə/ or /ɸ/): /gardə kot/, /pɔrtə kle/.

Rule 17: Remove /ɸ/ in final position of polysyllabic word (counting /ɸ/): /œ̃ mœbl ɑ̃sjɛ̃/, /il parl dɸ tu/, /œ̃ pɔrt mɔnɛ/, /œ̃ gard barjɛr/, /mil œ̃/, /dis kɔ̃tr œ̃/, /ɑ̃tr œ̃ e sɛ̃k/, /ɛl zɔ̃/. This rule, therefore, does not apply to the monosyllabic words ce, de, je, le, me, ne, etc.

However, /ə/ is kept before /œ̃/ in acte un, chapitre un, chiffre un, livre un, paragraphe un, scène un, strophe un, tome un, volume un; those combinations are in the lexicon.

Rule 18: Determine whether there are any /ɸ/ left in the breath-group. If so, starting from the left, change the first /ɸ/ to /ə/ if it follows two consonants or /j/ + consonant, remove it if otherwise; determine whether there are any remaining /ɸ/ in this modified breath-group; if so, apply same rule; and so on until all /ɸ/ have either been changed to /ə/ or removed: /nu parlɔ̃ d tu mɛ̃tnɑ̃/, /il nə l rədmɑ̃d pɑ/, /prane yn kɥijre/.

H. RESULTS OF FINAL TEST

Thousands of French sentences were used to develop and test all the rules which we presented in the previous chapters; in order to solve the many problems presented by a French text-to-speech program, we worked with sentences which had unusual levels of difficulty. As a final test for this research project, we tried those rules with the kinds of sentences an average user is likely to encounter.

For this final test, we selected 50 publications at random; in each of those, we chose one word at random and we took for our test the string of words, delimited by true punctuation marks, which contained that word. None of the potential problems which can be posed by punctuation marks (pages 5-9) were encountered. Thus, some of those test items are very

short (four words in strings 15 and 44) and some are very long (42 words in string 7).

The 50 test items are listed below:

1. bien accordé au vert-de-mer de ses yeux
   Colette, L'enfant malade
2. lesquelles les unes et les autres ne nous donnent que des renseignements insuffisants et d'ailleurs contradictoires
   Proust, A la recherche du temps perdu
3. Il donnait en exemple les oiseaux qui se mettent la tête sous l'aile
   Gide, Les Faux-Monnayeurs
4. C'est par la télévision et les radios qu'ils ont appris qu'il leur fallait se rendre dans l'établissement public le plus proche ou dans leur propre lycée
   Le Journal du Dimanche, June 12, 1983
5. et dont le titre était tel que je restai une minute immobile au-dessus comme sur un miroir
   Giraudoux, Suzanne et le Pacifique
6. l'acceptation de son frère modifiait ses dispositions
   Roger Martin du Gard, Les Thibault III
7. Ces arrestations sont intervenues au lendemain de la mise en état d'alerte maximum des forces de sécurité de l'Etat dans plusieurs grandes villes d'Espagne devant la crainte d'un «grave attentat» que pourrait commettre l'organisation indépendante basque E.t.a.
   La Dépêche du Midi, July 3, 1983
8. qu'il vient d'y entrer comme dans une confrérie enchantée
   Aragon, Les Voyageurs de l'Impériale
9. mais craignit d'affoler le paysan
   Malraux, L'Espoir
10. nous ne nous accordons que quelques instants de lucidité
    Colette, Au Milieu des Ruines
11. à l'heure où il sortait pour nettoyer la lampe de l'enseigne
    Duhamel, Nouvelle Rencontre de Salavin
12. Et je me rappelle que le décor de mon voyage intérieur était la terre du square
    Lacretelle, Le Christ aux bras étroits
13. Plus par nécessité que par gourmandise
    Giono, Que ma joie demeure
14. Les réactions au compromis conclu sur les montants compensatoires sont en général favorables
    Sud-Ouest, March 8, 1979
15. la tête la première
    Maupassant, Un soir
16. il semblait demander quelque chose avec insistance
    Mauriac, Le démon de la connaissance
17. comme le petit legs laissé par un oncle dont on pouvait espérer tout l'héritage
    Bazin, Il n'arrive jamais rien
18. brilla de nouveau d'un vif éclat jusqu'au soir
    Camus, Jonas
19. en quelques millions d'années
    Dutourd, Emile Trouche ou le diable et l'athée

20. une course disputée en deux parties
L'Indépendant (Perpignan), March 4, 1979
21. Il faut que nous obtenions la recette de la potion du druide
Goscinny & Uderzo, Astérix le Gaulois
22. Lorsqu'en 1945 fut créé en France le régime général de la Sécurité Sociale
Entre les lignes (Journal d'information de la RATP), Feb. 1983
23. ou plus libres dans une maison
50 millions de consommateurs, June 1983
24. C'est toujours quinze jours de gagnés
L'Express, June 2, 1983
25. il se heurtait contre les pierres
Mérimée, Colomba
26. qu'est-ce qui se passait donc là-dedans
Zola, L'Assommoir
27. Il n'avait pas la noire méchanceté de Domitien
Renan, Histoire des origines du Christianisme
28. au motif qu'elle est susceptible de troubler l'ordre public
France-Soir, June 13, 1983
29. Les gouvernants actuels s'efforcent de faire apparaître les opposants comme des factieux s'opposant à la loi de la République
Le Figaro, June 9, 1983
30. Cette aide s'adresse exclusivement aux propriétaires ou copropriétaires occupant eux-mêmes les lieux à titre de résidence principale
Ville de Paris (published by Hôtel de Ville, Paris), June 1983
31. L'intervention française dans ce domaine revêt des formes différentes selon le mode de coopération souhaité
Messages des PTT, March 1984
32. C'est leur reconnaître implicitement un statut de sapeurs et de casseurs de solidarité officiels
Le Canard Enchaîné, April 25, 1984
33. mais avec un tel scrupule
Bernanos, Sous le soleil de Satan
34. Qu'est-ce que je ferai de mes journées
Sartre, La Nausée
35. Les mouvements du trottoir et de la chaussée se troublaient de plus en plus
Romains, Mort de quelqu'un
36. mais de ceux dont les pieds toute la veille ont buté contre les escaliers
Giraudoux, Suzanne et le Pacifique
37. le rassemblement des corneilles dans la prairie de l'étang
Chateaubriand, Mémoires d'Outre-Tombe
38. mais leur costume où régnaient la fantaisie du goût individuel et le juste sentiment de la couleur
Gautier, Histoire du Romantisme
39. Ils n'avaient garde d'oublier ce caractère divin que l'antiquité avait attaché aux rois-pontifes et aux prêtres-fondateurs
Fustel de Coulanges, La Cité antique
40. c'est tout juste si j'arrivais à lacer mes chaussures
La Voix du Combattant, May 1984
41. donnant ainsi une économie de quatre livres environ
Poe/Baudelaire, Le canard au ballon

42. Deux animaux sont particulièrement recherchés
Bossu, Nouveaux voyages en Louisiane
43. D'autres arrivent à Paris avec des visions des Folies-Bergère
Daninos, Les carnets du Major Thompson
44. sont de qualité incomparable
Montagné, Larousse gastronomique
45. on répond que seule la rigueur est aimable
Le Canard Enchaîné, March 7, 1984
46. Au premier bruit d'un mal si étrange
Bossuet, Oraison funèbre de Henriette-Anne d'Angleterre
47. il lui sera délivré au vu des pièces justificatives un billet portant la mention «gratuit» ou établi avec une réduction de 50%
Ouest-France, June 21, 1983
48. Nous devînmes intimes par le corps et par l'esprit et nous allions ensemble nous promener en ville quelques heures par semaine
Céline, Voyage au bout de la nuit
49. c'est ainsi que des peuplades sauvages apprennent certaines nouvelles plusieurs jours avant que la poste les ait apportées à la colonie européenne
Proust, A la recherche du temps perdu
50. Une cape détourna le taureau
Montherlant, Les Bestiaires

A. The first step in our program (page 5) is to separate the text into input strings with a maximum of 30 words. Item 7 is the only one which contains more than 30 words; our sub-program is designed to split such strings at a medium point by using a list of prepositions and conjunctions. In this particular case, the division occurs after word 21 before dans.

B. The second step (page 9) is to check that no word is over thirty characters long; we find that the longest word (particulièrement in item 42) has only 16 characters.

C. The third step (page 10) is to rewrite the apostrophes.

D. The fourth step (page 11) is to rewrite the mathematical characters. In this group of 50 test items, there are only two numbers (1945 in item 22 and 50 in item 47) which need to be rewritten.

E. In the fifth step (page 13), we only need to rewrite % as poursan.

F. In the sixth step (page 13), we must take care of the end-of-line hyphens. The test items, as printed in the publications we used, contained 12 end-of-line hyphens (the rules are in Appendix III, pages 167-174):

établisse-ment in item 4	treated by rule 6
réac-tions in item 14	treated by rule 9b
sus-ceptible in item 28	treated by rule 9b
im-plicitement in item 32	treated by rule 9b
senti-ment in item 38	treated by rule 6
arri-vais in item 40	treated by rule 6
en-viron in item 41	treated by rule 9b
particulière-ment in item 42	treated by rule 6

Folies-Bergère in item 43	treated by rule 2c
ri-gueur in item 45	treated by rule 6
dé-livré in item 47	treated by rule 1
justificati-ves in item 47	treated by rule 6

G. In the seventh step (page 13), we find only one abbreviation (E.t.a. in item 7); our rules (page 144) yield the pronunciation /etea/.

H. In the eighth step (page 18), the features "cap" and "prnn" are set and all remaining capital letters are replaced with lower-case letters.

I. In the next step, the 627 words of the 50 test items go through the transcription and transformation programs.

There are 15 ambiguous spellings: (9 est, 4 plus, 1 donc, and 1 y); all the occurrences of est are such (c'est, elle est, est-ce) that they are found in the idiom list; two of the four occurrences of plus are also found in the idiom list. The 623 words (627 less donc, y, and the two remaining plus) which go through the transcription-transformation process are all transcribed correctly; the accuracy level is 100%.

J. Upon completing the transcription-transformation process, we find that 401 words out of 627 (64%) have been placed into an explicit grammar category, that 144 (23%) have been placed into an ambiguous category, and that 82 (13%) have not been placed into any category.

At this point, we refine the linguistic analysis by the method described on pages 52-77; except for the 5 cases discussed below, this additional data allows us to resolve the problems posed by the remaining ambiguous spellings (donc, y, and two plus), to divide the 32 items which have to be split into breath-groups shorter than fifteen syllables, to give the correct pronunciation to cinq in item 22, to add the consonants of liaison, and to drop the unstable vowels which can be removed.

The only comments which need to be made are:

Item 1: We cannot perform the semantic analysis which is necessary to determine whether bien in bien accordé is a noun (the meaning being possession granted) or an adverb (the meaning being well attuned). Since we cannot ascertain that bien is an adverb, we do not add the /n/ of liaison and our transcription is /bjɛ̃akɔrde/ instead of /bjɛ̃nakɔrde/.

Item 11: It would be preferable to cut after sortait, thus creating a 7-syllable group and an 11-syllable group; our program cuts after nettoyer, thus creating an 11-syllable group and a 7-syllable group.

Item 22: It would be preferable to cut after 1945, France, and général, thus creating groups having 10, 6, 7, and 9 syllables. Our program cuts after 1945 and général, thus creating only three groups with 10, 13, and 9 syllables. The absence of a cut after France is quite unnatural, but at the present time our program is not able to determine that créé is followed by a subject rather than by an object and our rules fail to divide the 32 syllables of item 22 into 10 and 22 syllables, then

into 10, 6, and 16 syllables, and finally into 10, 6, 7, and 9 syllables.

Item 31: In that item, we did not have sufficient data to identify the grammar categories of domaine and revêt and we placed the breath-group cut after revêt; cutting after domaine would be preferable.

Item 48: In that item, we placed the breath-group cuts after esprit and promener; cutting after intimes, esprit, ensemble, and ville would be preferable.

L E X I C O N

GENERAL COMMENTS 85

INDIVIDUAL WORDS

Group One: Words which are exceptions to the classifications set up by the table of endings 87

Group Two: Words which facilitate the assignment of linguistic features 111

Group Three: Words not transcribed correctly by our transcription rules 122

Group Four: Proper names not transcribed correctly by our transcription rules 132

Group Five: Spellings which have more than one pronunciation 137

Group Six: Abbreviations which have the "ww" marker and which are not pronounced letter by letter 142

Group Seven: Abbreviations which have the "ww" marker and in which the letter E is pronounced /e/ 144

Group Eight: Abbreviations which occur without periods and with small letters and which are not transcribed correctly by our transcription rules 144

IDIOMS 144

GENERAL COMMENTS

1. The rewritings described on pages 10-19 have already taken place when our computer program reaches the lexicon. Therefore, this lexicon shows aujourdui, quelkun, niy est, guesha, égosome, etc. instead of aujourd'hui, quelqu'un, n'est, ghesha, ægosome, etc.

In our computer program, all capital letters are replaced with lower-case letters; the linguistic features "cap" and "prnn" are set as explained on page 18. (For ease of reading, in this printed lexicon, we capitalize the words which can be only proper names.)

2. The reasons for including or not including in the lexicon some prefixed words and some words beginning with one of our 228 roots have been discussed on pages 22 and 25-29.

3. Since the table of endings changes ée, ie, and ue to é, i, and u, feminine forms are listed only when necessary to ensure correct linguistic features; thus, we have duré and durée, sorti and sortie, but only igné and ennemi. (The gender features are set only when the masculine and feminine forms are listed.) Words which have internal problems and for which the feminine is other than é(e), i(e), u(e) must appear in the lexicon with the masculine and the feminine forms (e.g., vosgien/vosgienne, bruxellois/bruxelloise, caennais/caennaise, adéquat/adéquate, benjamin/benjamine) unless —of course— only one form is an exception word (e.g., abrupt/abrupte).

4. If the singular form of a word is placed in the lexicon because of an internal problem, we must also list the plural form if it does not have the same pronunciation as the singular form (e.g., biennal/biennaux). (With oeuf/oeufs, boeuf/boeufs, the internal problem is only in the plural forms; the singular forms are transcribed correctly by our rules.)

5. The form without s and the form with s must be listed if they do not belong exactly to the same grammar category. Examples:

bien/biens	ce/ces	font/fonts	de/des
ensemble/ensembles	me/mes	étai/étais	fait/faits
contre/contres	la/las	lit/lits	ma/mas
entre/entres	moi/mois	part/parts	plu/plus
personne/personnes	soi/sois	pu/pus	ri/ris
rire/rires	sa/sas	se/ses	si/sis
point/points	son/sons	sort/sorts	souri/souris
suivant/suivants	su/sus	ta/tas	te/tes
ton/tons	tu/tus	être/êtres	vivre/vivres

6. The table of endings is an essential part of our program, but we must make sure that the rewritings performed by the table do not interfere with the transcription of words which have internal problems and/or that they do not assign wrong grammar categories.

Endings which appear without a final s do not pose a problem. For example, even though the endings end, rut, and um are on the table, it

is sufficient to list the singular form of the exception words zend, brut, and parfum. Problems arise only with the endings which are listed with and without a final s:

ents/ent: We must list thee singular and the plural forms of the words which have internal problems: chiendent/chiendents, équipotent/équipotents, quotient/quotients, pschent/pschents, etc. For example, if only the singular chiendent was listed, the plural form chiendents would not be found since the table of endings changes it to chiendan; if we listed only the form chiendan, the singular form would not be found since the table of endings would then assume it is a verb and would change it to chiende.

ais/ai: The words which are not verbs and which have a singular and a plural form must appear in the lexicon with both forms. For example, if vrai was the only form listed, vrais would be changed to vrai and would be found in the lexicon, but it would be marked "verb" since an explicit linguistic assignment made by the table of endings cannot be modified by the lexicon (see also soient, page 185).

ations/ation: The words which have internal problems (e.g., équation/s) could be listed with their singular and plural forms. However, since both endings assign the same explicit grammar category and since they are rewritten in the same way (assion), it is easier to enter them only in the singular rewritten form (équassion). We do likewise with ction/s and ption/s; thus, vivisecssion and tussuscepssion catch the singular and plural forms of vivisection and intussusception (after in is removed).

tions/tion: The words ending with tions which are not verbs and which do not have internal problems (e.g., ablutions) need be listed only in the plural since the table of endings will provide a correct transcription and correct features for the singular form. Both the singular and plural forms must be listed when the words have internal problems (e.g., équipartition/s, reddition/s).

ons/on: Since both endings are rewritten in the same way (on) and since the ending ons assigns an ambiguous feature ("nnvb") which —if necessary— can be corrected by the lexicon, it is sufficient to place the singular forms in the lexicon; this is true for the words which are not nouns (e.g., folichon) and for the words which have individual problems (e.g., charleston, paon).

iers/ier: The words which have internal problems (e.g., joaillier/s, quincaillier/s) could be listed with their singular and plural forms. However, since both endings assign the same explicit grammar category and since they are rewritten in the same way (ié), it is easier to enter them only in the singular rewritten form (e.g., joaillié, quincaillié). [However, since the table of endings would rewrite courrier (which has one /r/) and courriez (which has two /r/) in the same way (i.e., courrié), we have placed courrier and courriers in the lexicon.]

ers/er (except iers/ier): We must list the singular and the plural forms of the words which have internal problems: amer/s, cacaoyer/s, leader/s, etc. For example, if only the singular form cacaoyer was listed,

the plural form cacaoyers would not be found since the table of endings changes it to cacaoyé; if we listed only the form cacaoyé, both the singular and plural form would be transcribed correctly, but the singular form would be marked "infinitive".

irs/ir: The words which have internal problems need be listed only in the singular (e.g., menhir, septemvir) since the table of endings changes irs to ir and assigns the correct explicit grammar category.

7. We must ensure that words which exist only with a final s will not be assigned the wrong grammar category after the final s has been removed. Examples: bras, meurs, viens, éteins.

8. Words which are only proper names and which would not be transcribed correctly by our rules (e.g., Nevers) are in Group Four; when a word from Group Four occurs in the input string, the word is transcribed and the feature "prnn" is set (if not previously set in Step 8, page 18). Words which belong to the category of proper names and to some other category and which would be transcribed incorrectly by our rules are listed in Group One (e.g., clément/Clément), Group Two (e.g., gentil/Gentil), Group Three (e.g., adonis/Adonis), Group Five (e.g., eu/Eu), and the idiom list (e.g., sens/Sens). Words which are only proper names and which can be transcribed correctly by our rules (e.g., Alger) are not in the lexicon. The problems met in trying to assign the feature "prnn" are discussed on page 187.

INDIVIDUAL WORDS

GROUP ONE

This group contains the words which do not have the same linguistic features as those listed in the table of endings for a given ending and/or which cannot accept the rewritten ending because it would yield a wrong transcription.

Except for the transcription of a and â (see page 196), we use the same transcriptions as those recommended by the Petit Robert (1979).

Out of the 95 endings treated by the table of endings (see page 193), 23 endings (e.g., gents, ents, oiement) have no exceptions. The total of lexicon entries for the 72 endings which have exceptions is 1819.

The 1819 exceptions are listed below; the numbers 2, 4, 5, 6, 9, etc. correspond to the position of the 72 endings on the table.

2-Ending cents changed to çan and marked adjective, masc., plural.

Word which is not marked adjective:

cents nb,z

This word must be listed in order to prevent prefix separation:

décents adj,m,pl

4-Ending ément changed to éman and marked adverb.

Words which are not adverbs:

agrément	nn,m,sg	elément	nn,m,sg
cément	nn,m,sg	excrément	nn,m,sg
clément	adj,m,sg	récrément	nn,m,sg
complément	nn,m,sg	supplément	nn,m,sg
dément	adjnnvb	véhément	adj,m,sg
élément	nn,m,sg		

5-Ending emment changed to eaman and marked adverb.

The following word is a plural verb:

gemment

6-Ending amment changed to aman and marked adverb.

The following words are plural verbs:

enflamment	programment	renflamment

9-Ending alement changed to alman and marked adverb.

The following words are masculine singular nouns:

brimbalement	empalement	etalement	trimbalement
chevalement	enchevalement	ravalement	
dessalement	étalement	signalement	

10-Ending ellement changed to elman and marked adverb.

The following words are masculine singular nouns:

attellement	descellement	etincellement	nivellement
bossellement	encorbellement	grommellement	renouvellement
chancellement	ensorcellement	morcellement	ruissellement
cuvellement	étincellement	musellement	scellement

11-Ending eusement changed to euzman and marked adverb.

The following word is a masculine singular noun:

creusement

12-Ending gement changed to jman and marked noun, masculine, singular.

The following words are adverbs:

étrangement	etrangement	sagement	sauvagement

13-Ending onnement changed to onnman and marked noun, masc., singular.

The following words are adverbs:

bonnement gloutonnement

14-Ending ssement changed to sman and marked noun, masculine, singular.

The following words are adverbs:

bassement faussement grassement

15-Ending chement changed to cheman and marked noun, masc., sing.

The following words are adverbs:

chichemen	franchement	lâchement	sèchement
fraîchement	gauchement	richement	vachement

16-Ending ivement changed to ivman and marked adverb.

The following word is a masculine singular noun:

enjolivement

18-Ending scent changed to ssan and marked noun, masculine, singular.

The following words are plural verbs:

acquiescent immiscent

19-Ending guent changed to gue and marked verb, plural.

The following word is a masculine singular noun:

onguent

20-Remaining ending uent changed to u and marked verb, plural.

Words which are not verbs:

congruent	adj,m,sg	eloquent	adj,m,sg
conséquent	adj,m,sg	fréquent	adj,m,sg
défluent	nn,m,sg	subséquent	adj,m,sg
éloquent	adj,m,sg		

These three words appear on the list of ambiguous spellings:

affluent confluent influent

21-Ending ient changed to i and marked verb, plural.

The following words are singular verbs:

abstient advient appartient circonvient

contrevient	obtient	soutient	vient
devient	parvient	souvient	
disconvient	provient	survient	
maintient	ressouvient	tient	

Words which are not verbs:

client	nn,m,sg	inconvénient	nn,m,sg
coefficient	nn,m,sg	ingrédient	nn,m,sg
conscient	adj,m,sg	omniscient	adj,m,sg
efficient	adj,m,sg	orient	nn,m,sg
émollient	nn,m,sg	patient	adj,m,sg
emollient	nn,m,sg	prescient	adj,m,sg
escient	nn,m,sg	quotient	nn,m,sg
impatient	adj,m,sg	récipient	nn,m,sg

These three words appear on the list of ambiguous spellings:

convient	expédient	obvient

22-Remaining ending ent changed to e and marked verb, plural.

The following words are singular verbs:

ment	repent	ressent	sent

Words which are not verbs:

abrivent	nn,m,sg	bâtiment	nn,m,sg
absent	adjnn,m,sg	blanchiment	nn,m,sg
absolument	adv	boniment	nn,m,sg
abstinent	adj,m,sg	calament	nn,m,sg
accident	nn,m,sg	châtiment	nn,m,sg
adhérent	nn,m,sg	chiendent	nn,m,sg
adjacent	adj,m,sg	ciment	nn,m,sg
afférent	adj,m,sg	cohérent	adj,m,sg
agent	nn,m,sg	compartiment	nn,m,sg
aliment	nn,m,sg	compétent	adj,m,sg
ambigument	adv	compliment	nn,m,sg
antécédent	nn,m,sg	concurrent	nn,m,sg
antiferment	nn,m,sg	condiment	nn,m,sg
apparent	adj,m,sg	confident	nn,m,sg
ardent	adj,m,sg	constringent	adj,m,sg
argent	nn,m,sg	continent	nn,m,sg
argument	nn,m,sg	contingent	nn,m,sg
arpent	nn,m,sg	continûment	adv
assentiment	nn,m,sg	contrevent	nn,m,sg
assidûment	adv	corpulent	adj,m,sg
assortiment	nn,m,sg	crûment	adv
astringent	nn,m,sg	décadent	adj,m,sg
augment	nn,m,sg	décent	adj,m,sg
auvent	nn,m,sg	déférent	adj,m,sg
avent	nn,m,sg	dénoûment	nn,m,sg

dent	nn,f,sg
déponent	adj,m,sg
détriment	nn,m,sg
différent	adj,m,sg
diligent	adj,m,sg
dissentiment	nn,m,sg
dissident	nn,m,sg
dissolument	adv
document	nn,m,sg
dolent	adj,m,sg
dûment	adv
éminent	adj,m,sg
eminent	adj,m,sg
émolument	nn,m,sg
emolument	nn,m,sg
engoulevent	nn,m,sg
entregent	nn,m,sg
éperdument	adv
eperdument	adv
équipollent	adj,m,sg
equipollent	adj,m,sg
équipotent	adj,m,sg
equipotent	adj,m,sg
étourdiment	adv
etourdiment	adv
évent	nn,m,sg
event	nn,m,sg
excédent	nn,m,sg
féculent	nn,m,sg
fervent	adj,m,sg
fichûment	adv
filament	nn,m,sg
firmament	nn,m,sg
flatulent	adj,m,sg
fourniment	nn,m,sg
fragment	nn,m,sg
froment	nn,m,sg
gent	nn,f,sg
gentiment	adv
goulûment	adv
grûment	adv
hardiment	adv
immanent	adj,m,sg
imminent	adj,m,sg
impénitent	adj,m,sg
impertinent	nn,m,sg
impotent	nn,m,sg
imprudent	adjnn,m,sg
impudent	adj,m,sg
incident	nn,m,sg
indéfiniment	adv
indigent	nn,m,sg
indulgent	adj,m,sg
infiniment	adv
ingénument	adv
inhérent	adj,m,sg
innocent	adjnn,m,sg
instrument	nn,m,sg
intelligent	adj,m,sg
intermittent	adj,m,sg
irrésolument	adv
jacent	adj,m,sg
joliment	adv
jument	nn,f,sg
latent	adj,m,sg
lent	adj,m,sg
ligament	nn,m,sg
linéament	nn,m,sg
liniment	nn,m,sg
mécontent	adj,m,sg
médicament	nn,m,sg
moment	nn,m,sg
monovalent	adj,m,sg
monument	nn,m,sg
nûment	adv
occident	nn,m,sg
occurrent	adj,m,sg
omnipotent	adj,m,sg
opulent	adj,m,sg
orpiment	nn,m,sg
patent	adj,m,sg
pénitent	nn,m,sg
permanent	adjnn,m,sg
pertinent	adj,m,sg
pigment	nn,m,sg
piment	nn,m,sg
plurivalent	adj,m,sg
poliment	adv
polyvalent	adjnn,m,sg
précédent	adjnn,m,sg
présent	adjnn,m,sg
pressentiment	nn,m,sg
proéminent	adj,m,sg
prudent	adj,m,sg
pschent	nn,m,sg
pulvérulent	adj,m,sg
purulent	adj,m,sg
quasiment	adv
rassortiment	nn,m,sg
réassortiment	nn,m,sg
récent	adj,m,sg
récurrent	adj,m,sg
réfringent	adj,m,sg
régent	nn,m,sg
régiment	nn,m,sg
relent	nn,m,sg

rémanent	adj,m,sg	talent	nn,m,sg
rémittent	adj,m,sg	tangent	adj,m,sg
résolument	adv	tégument	nn,m,sg
ressentiment	nn,m,sg	tempérament	nn,m,sg
restringent	nn,m,sg	testament	nn,m,sg
rudiment	nn,m,sg	torrent	nn,m,sg
sanguinolent	adj,m,sg	tourment	nn,m,sg
sarment	nn,m,sg	transparent	adj,m,sg
sédiment	nn,m,sg	trident	nn,m,sg
segment	nn,m,sg	truculent	adj,m,sg
sentiment	nn,m,sg	turbulent	adj,m,sg
sergent	nn,m,sg	uniment	adv
serment	nn,m,sg	univalent	adj,m,sg
serpent	nn,m,sg	urgent	adj,m,sg
souvent	adv	vent	nn,m,sg
strident	adj,m,sg	ventripotent	adj,m,sg
succulent	adj,m,sg	virulent	adj,m,sg
suréminent	adj,m,sg	vraiment	adv

These nineteen words appear on the list of ambiguous spellings:

coïncident	content	convergent	couvent	détergent
divergent	émergent	équivalent	évident	excellent
ferment	insolent	négligent	parent	président
pressent	résident	somnolent	violent	

The words accent, cent, and comment appear on the idiom list.

23-Ending ais changed to ai and marked verb, singular.

Words which are not solely verbs:

ais	nn,m,an	écossais	adjnn,m,an
albanais	adjnn,m,an	ecossais	adjnn,m,an
anglais	adjnn,m,an	engrais	nn,m,an
antillais	adjnn,m,an	épais	adj,m,an
aragonnais	adjnn,m,an	epais	adj,m,an
bâbordais	nn,m,an	essais	nn,m,pl
bais	adj,pl	étais	nnvb
balais	nn,m,pl	etais	nnvb
bayonnais	adjnn,m,an	frais	adjnn,m,an
biais	nn,m,an	français	adjnn,m,an
brais	nn,m,pl	gais	adj,m,pl
caennais	adjnn,m,an	geais	nn,m,pl
chais	nn,m,pl	groenlandais	adjnn,m,an
confolennais	adjnn,m,an	harnais	nn,m,an
craonais	adjnn,m,an	havrais	adjnn,m,an
dadais	nn,m,an	hollandais	adjnn,m,an
dais	nn,m,an	irlandais	adjnn,m,an
déblais	nn,m,pl	jais	nn,m,an
délais	nn,m,pl	jamais	adv,z;83
désormais	adv	japonais	adjnn,m,an
doullennais	adjnn,m,an	lais	adjnn,pl

landais	adjnn,m,an	papegais	nn,m,pl
laquais	nn,m,an	piémontais	adjnn,m,an
liais	nn,m,an	polonais	adjnn,m,an
mais	coord,z	portugais	adjnn,m,an
malais	adjnn,m,an	punais	adj,m,an
maltais	adjnn,m,an	quais	nn,m,pl
marais	nn,m,an	rabais	nn,m,an
mauvais	adj,m,an	rais	nn,m,pl
milanais	adjnn,m,an	relais	nn,m,an
minerais	nnvb	remblais	nn,m,pl
niais	adj,m,an	rouennais	adjnn,m,an
ouais	adv	tribordais	nn,m,an
palais	nn,m,an	virelais	nn,m,pl
panais	nn,m,an	vrais	adj,pl

24-Ending ai changed to é and marked verb, singular.

(This rewriting cannot result in transcribing c and g as /s/ and /ʒ/ instead of /k/ and /g/ since there are no verbs ending with cai or gai.)

Words which are not solely verbs:
(balai, essai, etc. are marked "nnvb" since, for example, balaies becomes balaie when final s is removed, and then balai in the table of endings.)

bai	adjnn,sg	gai	adj,sg
balai	nnvb	geai	nn,m,sg
brai	nnvb	lai	adjnn,sg
chai	nn,m,sg	mai	nn,m,sg
déblai	nnvb	papegai	nn,m,sg
délai	nnvb	quai	nn,m,sg
essai	nnvb	rai	nnvb
étai	nnvb	remblai	nnvb
etai	nnvb	virelai	nn,m,sg
frai	nnvb	vrai	adj,sg

The word minerai is on the list of ambiguous spellings.

25-Ending ez changed to é and marked verb, plural.

Words which are not verbs:

chez	prep,z;97,102,112	merguez	nn,f,an
lez	prep	nez	nn,m,an
fez	nn,m,an		

The words Suez and Rodez are on the list of ambiguous spellings and the word assez is on the list of idioms.

26-Ending ix changed to i and marked noun, ambig. gender, ambig. number.

The numbers dix and six are on the list of idioms.

These nouns (masculine, ambiguous number) must not be changed to i

because the letter x is pronounced:

aptérix	phoenix	strix
hélix	préfix	tamarix
phénix	sandix	trionix

27-Ending ux changed to u and marked adjective/noun.

Words which are not marked adjectives/nouns:

deux	nb,z	mieux	nnadv,z
équivaux	v,sg	peux	v,sg
equivaux	v,sg	prévaux	v,sg
eux	allpron,m,pl;98	vaux	nnvb
faux	adjnadvb	veux	v,sg

The word ceux is on the idiom list.

Word in which ux must not be changed to u because x is pronounced:

lux	nn,m,an

The words which are specifically nouns and which have a final x in the singular are listed in order to facilitate the linguistic analysis as in, for example, taux anglais:

chaux	nn,f,an	houx	nn,m,an
courroux	nn,m,an	queux	nn,m,an
époux	nn,m,an	reflux	nn,m,an
epoux	nn,m,an	saindoux	nn,m,an
flux	nn,m,an	toux	nn,f,an
gueux	nn,m,an		

28-Ending ations changed to assion and marked noun, feminine, plural.

Words which are not feminine nouns:

acclimations	v,pl	épations	v,pl
calfations	v,pl	epations	v,pl
casemations	v,pl	érations	v,pl
cations	nn,m,pl	erations	v,pl
colmations	v,pl	frelations	v,pl
constations	v,pl	himations	nn,m,pl
cravations	v,pl	mandations	v,pl
dations	v,pl	mations	v,pl
dérations	v,pl	ouations	v,pl
dilations	v,pl	pirations	v,pl
éclations	v,pl	sulfations	v,pl
eclations	v,pl		

These three words appear on the list of ambiguous spellings:

rations	relations	translations

29-Ending ctions changed to cssion and marked noun, feminine, plural.

The following words are plural verbs:

édictions	délections	respections
edictions	humections	suspections

These thirteen words appear on the list of ambiguous spellings:

affections	concoctions	contractions	détractions
dictions	diffractions	infections	injections
inspections	interjections	objections	réfractions
rétractions			

30-Ending ptions changed to pssion and marked noun, feminine, plural.

The following words are plural verbs:

adaptions	cooptions	escomptions
captions	décryptions	mécomptions
comptions	domptions	sculptions

These six words appear on the list of ambiguous spellings:

acceptions	adoptions	exceptions	exemptions
interceptions	options		

31-Remaining ending tions changed to tion and marked verb, plural.

The following words are feminine plural nouns:

ablutions	coalitions	dilutions	exhibitions
abolitions	coercitions	diminutions	expéditions
absolutions	combustions	discrétions	expositions
abstentions	commotions	disparitions	extraditions
acquisitions	comparutions	dispositions	finitions
additions	compétitions	disproportions	gestions
admixtions	compositions	disquisitions	hypersécrétions
admonitions	concrétions	dissolutions	hyposécrétions
allocutions	conditions	distributions	ignitions
ambitions	constitutions	ébullitions	immixtions
apparitions	contraventions	ebullitions	impositions
appétitions	contributions	élocutions	inanitions
appositions	contritions	elocutions	inattentions
assertions	conventions	émotions	inexécutions
attributions	déglutitions	emotions	inhibitions
attritions	démolitions	équipartitions	inquisitions
auditions	dentitions	equipartitions	insertions
bastions	destitutions	éruditions	institutions
cautions	détentions	eruditions	interventions
circonlocutions	dévolutions	évolutions	intuitions
circonventions	dévotions	evolutions	involutions
circonvolutions	digestions	exhaustions	invotions

irrésolutions	perditions	prostitutions	restitutions
juxtapositions	perquisitions	punitions	rétentions
locomotions	pétitions	questions	rétributions
locutions	pollutions	réapparitions	révolutions
lotions	positions	redditions	séditions
manutentions	potions	rédhibitions	solutions
mixtions	prénotions	réexpéditions	substitutions
monitions	présuppositions	réimpositions	subventions
motions	prétentions	réinsertions	suggestions
munitions	prétéritions	réparitions	sujétions
nutritions	préventions	répartitions	superstitions
obtentions	prohibitions	répétitions	suppositions
oppositions	promotions	réplétions	traditions
partitions	proportions	réquisitions	volitions
parturitions	propositions	résolutions	

These seventeen words appear on the list of ambiguous spellings:

attentions	contentions	désertions	éditions
électrocutions	excrétions	exécutions	intentions
inventions	mentions	notions	persécutions
portions	rééditions	réinventions	sécrétions
transitions			

34-Remaining ending tion changed to ssion and marked noun, fem., sing.

The following words are masculine singular nouns:

cation	himation

35-Remaining ending ons changed to on and marked noun/verb.

Word which is not a noun and which does not have a form without a final s:

abscons	adj,m,an

36-Remaining ending on unchanged and marked noun, singular.

Words which are not solely nouns:

bon	adj,m,sg	sinon	coord
folichon	adj,m,sg	son	detnn,sg,n;94,103
mon	det,ag,sg,n;103	ton	detnn,sg,n;94,103

The words non, on, selon are on the list of idioms.

In the following masculine singular nouns, on is not pronounced /ɔ/:

bacon	chatterton	faon	paon
badminton	eleison	klaxon	taon
charleston	epsilon	omicron	upsilon

37-Ending ers changed to é and marked noun, masculine, plural.

This word ends with /e/, but it is not a noun:

volontiers adv

The following masculine plural nouns must not be changed to é because the pronunciation is not /e/:

abers
amylobacters
asters
badinters
belvéders
bers
bessemers
bitters
blazers
blisters
bookmakers
boosters
bootleggers
boxers
breeders
bulldozers
bunkers
cancers
canters
carters
cathéters
challengers
charters
chesters
clippers
cockers
containers
corners
coroners
cutters
ders
destroyers
dockers
drivers
eiders
enfers
esters
éthers
ethers
feeders
fers
flippers
fraters
führers
gangsters
géasters
geysers
hamburgers
hamsters
hivers
impers
inters
interviewers
jiggers
jokers
kaisers
khmers
kippers
kreutzers
lasers
leaders
libers
lieders
mâchefers
magisters
managers
masers
masséters
messers
minnesängers
minnesingers
munsters
outriggers
outsiders
palmers
partners
paters
placers
pokers
polders
polyesters
porters
posters
quakers
racketters
reporters
retrievers
revolvers
scanners
schnauzers
schooners
scooters
setters
shakers
sitters
skippers
spalters
speakers
spencers
sphincters
spiders
spoilers
sprinters
squatters
stadhouders
starters
stathouders
steamers
strippers
subers
supers
supporters
sweaters
tankers
teenagers
tenders
thalers
trochanters
trochiters
ulsters
vétivers
vomers
waters
webers

Other words which must not be changed to é because the pronunciation is not /e/:

acquiers	v,sg	enquiers	v,sg
amers	adj,m,pl	envers	prep;97,102,112
avers	nn,m,an	ers	nn,m,an
chers	adj,m,pl	maters	nn,f,pl
convers	adj,m,an	obvers	nn,m,an
cuillers	nn,f,pl	pers	adj,m,an
devers	prep	pervers	adj,m,an
dévers	nn,m,an	quiers	v,sg
divers	adndp,m,pl,z	ressers	v,sg

revers	nn,m,an	univers	nn,m,an
sers	v,sg	vers	nnprep,z;97,102,112
travers	nn,m,an		

The words fiers, mers, and tiers are on the list of idioms.

38-Ending ier changed to ié and marked noun, masculine, singular.

The following words are infinitives:

acidifier	crier	identifier	planifier
affilier	crucifier	incendier	plier
allier	dédier	ingénier	préjudicier
amnistier	déifier	initier	prier
amodier	déprécier	injurier	privilégier
amplifier	dévier	interfolier	psalmodier
anémier	différencier	inventorier	publier
anesthésier	disgracier	irradier	purifier
apostasier	disqualifier	justifier	putréfier
apparier	dissocier	lénifier	qualifier
apprécier	diversifier	licencier	radier
approprier	domicilier	lier	radiographier
armorier	dulcifier	liquéfier	rallier
asphyxier	écrier	lithographier	ramifier
associer	ecrier	lubréfier	rapatrier
atrophier	édifier	magnifier	rapparier
aurifier	edifier	manier	rapproprier
autographier	enlier	marier	raréfier
autopsier	envier	méfier	rassasier
avarier	épier	mendier	ratifier
balbutier	epier	mésallier	razzier
barbifier	estropier	modifier	réconcilier
béatifier	étudier	mollifier	récrier
bénéficier	etudier	momifier	rectifier
bêtifier	excommunier	mortifier	réédifier
bonifier	excorier	mystifier	réexpédier
calligraphier	exfolier	négocier	réfugier
calomnier	expatrier	nier	remédier
carier	expédier	nitrifier	remercier
certifier	expier	notifier	remplier
charrier	exproprier	obvier	renvier
châtier	extasier	orthographier	répudier
clarifier	falsifier	ossifier	résilier
classifier	fortifier	oublier	rubéfier
cocufier	fructifier	pacifier	sacrifier
codifier	gabarier	pallier	salarier
colorier	gazéifier	panifier	sanctifier
communier	glorifier	parier	saponifier
concilier	gracier	parodier	scarifier
congédier	gratifier	personnifier	scier
contrarier	harmonier	pétrifier	sérier
convier	historier	photographier	signifier
copier	humilier	plagier	simplifier

solfier	stipendier	terrifier	vérifier
solidifier	stratifier	tonifier	versifier
soucier	stupéfier	torréfier	vicier
spécifier	supplicier	trier	vitrifier
spolier	supplier	tuméfier	vivifier
sténographier	télégraphier	varier	

Other words which are not masculine singular nouns:

dernier	adj,m,sg	officier	nninf
entier	adj,m,sg	premier	adj,m,sg
familier	adj,m,sg	séculier	adj,m,sg
grossier	adj,m,sg	singulier	adjnn,m,sg
hier	adv		

The word fier is on the list of idioms.

39-Remaining ending er changed to é and marked infinitive.

The following words have final /e/, but are not solely infinitives:

aller	nninf	gaucher	nn,m,sg
andouiller	nn,m,sg	goûter	inf,npref,nninf
archer	nn,m,sg	horloger	nn,m,sg
baiser	inf,npref,nninf	léger	adj,m,sg
berger	nn,m,sg	loyer	nn,m,sg
boucher	inf,npref,nninf	manger	inf,npref,nninf
boulanger	nninf	maraîcher	nn,m,sg
bouteiller	nn,m,sg	ménager	inf,npref,adjinf
bûcher	inf,npref,nninf	mensonger	adj,m,sg
cacaoyer	nn,m,sg	messager	nn,m,sg
caloyer	nn,m,sg	métayer	nn,m,sg
clocher	inf,npref,nninf	noyer	nninf
cocher	inf,npref,nninf	oreiller	nn,m,sg
conseiller	inf,npref,nninf	papayer	nn,m,sg
cornouiller	nn,m,sg	parler	nninf
coucher	inf,npref,nninf	passager	nn,m,sg
danger	nn,m,sg	pêcher	inf,npref,nninf
débucher	nninf	penser	inf,npref,nninf
déjeuner	nninf	plaidoyer	nn,m,sg
dîner	nninf	potager	nn,m,sg
écailler	nninf	poulailler	nn,m,sg
ecailler	nninf	pourparler	nn,m,sg
écuyer	nn,m,sg	rucher	nninf
ecuyer	nn,m,sg	souper	inf,npref,nninf
électroménager	nn,m,sg	toucher	inf,npref,nninf
electroménager	nn,m,sg	usager	nn,m,sg
étranger	nn,m,sg	vacher	nn,m,sg
etranger	nn,m,sg	verger	inf,npref,nn,m,sg
foyer	nn,m,sg	viager	adj,m,sg
fromager	nn,m,sg	voyer	inf,npref,nn,m,sg

The following masculine singular nouns do not have a final /e/:

aber	dogger	lieder	spalter
alter	driver	mâchefer	speaker
amylobacter	eider	magister	spencer
aster	enfer	maser	sphincter
badinter	éther	masséter	spider
belvéder	ether	messer	spoiler
ber	feeder	minnesänger	stadhouder
bessemer	fer	minnesinger	starter
blazer	frater	munster	stathouder
blister	führer	outrigger	steamer
bookmaker	gangster	outsider	suber
booster	géaster	partner	sweater
bootlegger	geiger	pater	tanker
breeder	geyser	poker	teenager
bulldozer	hamburger	polder	téléreporter
bunker	hamster	polyester	tender
cancer	hiver	quaker	thaler
canter	hyper	racketter	trochanter
cathéter	imper	radioreporter	trochiter
challenger	inter	retriever	ulster
charter	jigger	revolver	ver
chester	joker	scanner	vétiver
clipper	kaiser	schnauzer	vomer
cocker	khmer	schooner	water
container	kipper	scooter	weber
coroner	kreutzer	setter	welter
cutter	laser	shaker	
destroyer	leader	sitter	
docker	liber	skipper	

Other words which do not have a final /e/:

amer	adj,m,sg	der	nn,ag,sg
cascher	adj,ag,an	kascher	adj,ag,an
casher	adj,ag,an	kasher	adj,ag,an
cawcher	adj,ag,an	mer	nn,f,sg
cher	adj,m,sg	outremer	adjnn,ag,sg
confer	imp	quater	adv
cuiller	nn,f,sg	ter	adv

These nineteen words appear on the list of ambiguous spellings:

bitter	cacher	carter	corner	ester	flipper	interviewer
manager	mater	palmer	placer	porter	poster	reporter
sprinter	squatter	stripper	super	supporter		

40-Ending oir unchanged and marked noun, masculine, singular.

The following words are infinitives:

apercevoir	apparoir	asseoir	cevoir

chaloir	emouvoir	percevoir	seoir
choir	équivaloir	pleuvoir	surseoir
comparoir	equivaloir	pourvoir	valoir
échoir	falloir	promouvoir	voir
echoir	messeoir	rasseoir	vouloir
émouvoir	mouvoir	ravoir	

The following words are nouns and infinitives:

avoir	devoir	pouvoir	savoir

43-Remaining ending ir unchanged and marked infinitive.

The following words are masculine singular nouns:

avenir	elixir	plaisir	tir
casimir	émir	quindécimvir	triumvir
cuir	emir	sabir	vizir
désir	fakir	saphir	
duumvir	loisir	septemvir	
élixir	menhir	soupir	

The following words are nouns and infinitives:

devenir	repentir	tapir

44-Ending ndre unchanged and marked infinitive.

Words which are not solely infinitives:

bélandre	nn,f,sg	méandre	nn,m,sg
calandre	nnvb	moindre	adj,ag,sg
cendre	nnvb	monandre	adj,ag,sg
coriandre	nn,f,sg	oléandre	nn,m,sg
cylindre	nnvb	palissandre	nn,m,sg
effondre	v,sg	périchondre	nn,m,sg
engendre	v,sg	polyandre	adj,ag,sg
esclandre	nn,m,sg	salamandre	nn,f,sg
filandre	nn,f,sg	scaphandre	nn,m,sg
gendre	nn,m,sg	scolopendre	nn,f,sg
gindre	nn,m,sg	tendre	adjinf
hypocondre	adj,ag,sg	triandre	adj,ag,sg
malandre	nn,f,sg		

45-Ending ont changed to on and marked verb, plural.

Words which are not verbs (fonts must be included since it does not have a singular form):

affront	nn,m,sg	front	nn,m,sg
amont	nn,m,sg	mont	nn,m,sg
dont	sub,t;109,111	rodomont	adjnn,m,sg
fonts	nn,m,pl		

The word pont is on the idiom list.

46-Ending end changed to ean and marked verb, singular.
(The change is made to ean rather than to an so that words ending with cend or gend will be transcribed correctly.)

The following words are masculine singular nouns:

différend révérend zend

47-Ending ra unchanged and marked verb, singular.

Words which are not solely verbs:
(Words which, like gras, do not have a form without s must also be included on the following list. Bras, semen-contra, and et cetera are on the idiom list.)

abracadabra	nn,m,sg	hourra	nn,m,sg
agora	nn,f,sg	hypocras	nn,m,an
angora	adj,ag,sg	jettatura	nn,f,sg
ara	nn,m,sg	libera	nn,m,an
aspidistra	nn,m,sg	madras	nn,m,an
baccara	nn,m,sg	matras	nn,m,an
bora	nn,f,sg	nomenklatura	nn,f,sg
chistera	nn,f,sg	nostras	adj,m,an
choléra	nn,m,sg	opéra	nnvb
cobra	nn,m,sg	para	nnvb
datura	nn,m,sg	patatras	nn,m,an
débarras	nn,m,an	phylloxera	nn,m,sg
dégras	nn,m,an	plâtras	nn,m,an
droséra	nn,f,sg	ra	nn,m,sg
embarras	nn,m,an	rémora	nn,m,sg
extra	adj,ag,sg	sassafras	nn,m,an
eyra	nn,m,sg	sierra	nn,f,sg
fatras	nn,m,an	sophora	nn,m,sg
fouchtra	adv	stras	nn,m,an
gandoura	nn,f,sg	tétras	nn,m,an
gras	adj,m,an	toundra	nn,f,sg
haras	nn,m,an	ultra	adj,ag,sg

The word ras is on the list of ambiguous spellings.

48-Ending ait changed to ai and marked verb, singular.

Words which are not solely verbs (the plural form is listed when it does not belong to the same grammatical category as the form without s):

abstrait	adjnvbpp	défait	adjnvbpp
abstraits	adjnnpp,m,pl	défaits	adjnnpp,m,pl
attrait	nn,m,sg	distrait	adjnvbpp
bienfait	nn,m,sg	distraits	adjnnpp,m,pl
contrefait	adjnvbpp	extrait	adjnvbpp
contrefaits	adjnnpp,m,pl	extraits	adjnnpp,m,pl

forfait	adjnvbpp	retrait	nn,m,sg
forfaits	adjnnpp,m,pl	satisfait	adjnvbpp
imparfait	nn,m,sg	satisfaits	adjnnpp,m,pl
lait	nn,m,sg	souhait	nn,m,sg
méfait	nn,m,sg	stupéfait	adjnvbpp
parfait	adjnvbpp	stupéfaits	adjnnpp,m,pl
parfaits	adjnnpp,m,pl	surfait	adjnvbpp
portrait	nn,m,sg	surfaits	adjnnpp,m,pl
refait	adjnvbpp	trait	adjnvbpp
refaits	adjnnpp,m,pl	traits	adjnnpp,m,pl

The words fait and faits are on the idiom list.

49-Ending rut changed to ru and marked verb, singular.

Words which are not verbs:

rut	nn,m,sg	brut	adj,m,sg

50-Ending ionne unchanged and marked verb, singular.

Words which are not solely verbs:

championne	nn,f,sg	lionne	nn,f,sg
espionne	nnvb		

51-Ending eur unchanged and marked noun, singular.

Words which are not nouns:
(The words ending with eurs must be in the lexicon since, otherwise, the table of endings would assign wrong features after final s is removed.)

ailleurs	adv	leurs	det,ag,pl,z;104
leur	dp,ag;79,87,88,92, 94,95,100,105,108	meurs	v,sg
		plusieurs	adndp,pl,z;94,95,96,104

Word in which final eur is not pronounced /œr/:

monsieur	nn,m,sg

52-Ending ique changed to ike and marked adjective, ambi. gender, sing.

Words which are not solely adjectives (dépiques, intoxiques, repiques are included in order to prevent prefix separation):

abdique	v,sg	communique	v,sg
alambique	v,sg	complique	v,sg
applique	nnvb	cortique	v,sg
astique	v,sg	critique	nnvb
authentique	adjvb	dépique	v,sg
brique	nnvb	dépiques	v,sg
chique	nnvb	diagnostique	v,sg
clique	nnvb	dique	v,sg

domestique	adjnnvb	plastique	adjnnvb
duplique	v,sg	polémique	nnvb
encaustique	nnvb	politique	adjnnvb
étrique	v,sg	pratique	adjnnvb
etrique	v,sg	prévarique	v,sg
explique	v,sg	pronostique	v,sg
fabrique	nnvb	rapplique	v,sg
fornique	v,sg	repique	v,sg
imbrique	v,sg	repiques	v,sg
implique	v,sg	réplique	nnvb
intoxique	v,sg	rustique	adjvb
intoxiques	v,sg	sophistique	v,sg
mastique	v,sg	syndique	v,sg
mimique	nnvb	tique	v,sg
musique	nnvb	trafique	v,sg
oblique	adjvb	trique	nnvb
panique	nnvb	vendique	v,sg
pique	nnvb		

The following adjectives must be in the lexicon since, otherwise, they would receive wrong linguistic features after prefix separation:

hydropique	adj,ag,sg	hypnotiques	adj,ag,pl
hydropiques	adj,ag,pl	thermoplastique	adj,ag,sg
hypnotique	adj,ag,sg	thermoplastiques	adj,ag,pl

53-Ending ic changed to ik and marked noun, masculine, singular.

Word which is not a masculine noun:

sic	adv

54-Ending ien unchanged and marked noun, masculine, singular.

The following words are singular verbs:
(The words ending with iens must be in the lexicon since, otherwise, the table of endings would assign wrong linguistic features after the removal of final s.)

abstiens	disconviens	proviens	surviens
appartiens	maintiens	ressouviens	tiens
contreviens	obtiens	soutiens	viens
deviens	parviens	souviens	

Other words which are not masculine nouns:

ancien	adj,m,sg	rien	nnad,n;83,106
biens	nn,m,pl	riens	nn,m,pl
combien	adv		

The word bien, the ambiguous string les tiens, and the combinations in which tiens is not a verb (e.g., aux tiens) are on the list of idioms.

55-Ending vienne changed to viène and marked verb, singular.

Words which are not verbs:

antédiluvienne	adj,f,sg	diluvienne	adj,f,sg
bolivienne	nn,f,sg	pelvienne	adj,f,sg
cracovienne	nn,f,sg	péruvienne	nn,f,sg

56-Ending tienne changed to tiène and marked verb, singular.

The following words are feminine nouns:
(The plural antiennes must be included to prevent prefix separation.)

antienne	chrétienne	laotienne	tahitienne
antiennes	égyptienne	lilliputienne	tribunitienne
béotienne	egyptienne	martienne	vénitienne
capétienne	haïtienne	rhétienne	

Other words which are not singular verbs:

kantienne	adj,f,sg	aoûtienne	adj,f,sg

The ambiguous strings la tienne, les tiennes and the combinations in which tiennes is not a verb (e.g., aux tiennes) are on the list of idioms.

58-Ending ante unchanged and marked adjective, feminine, singular.

The following words are singular verbs:
(The forms déplantes, désaimantes, replantes are included in order to prevent prefix separation; if prefix separation was allowed, these verbs would receive the same linguistic features as plante and aimante.)

cante	désaimantes	enfante	replante
chante	diamante	ensanglante	replantes
déplante	ébouillante	gante	supplante
déplantes	ebouillante	hante	vante
désaimante	enchante	implante	

Other words which are not solely adjectives:

aimante	adjvb	huitante	nb
andante	nnadv	nonante	nb
andantes	nn,m,pl	octante	nb
brillante	adjvb	pante	nn,m,sg
brocante	nnvb	plaisante	adj,f,sg,npref,adjvb
épouvante	nnvb	plante	nnvb
epouvante	nnvb	septante	nb
fainéante	adjnnvb		

The words quarante, cinquante, and soixante are on the list of idioms.

59-Ending ance unchanged and marked noun, feminine, singular.

Words which are not solely feminine nouns:

(The form décontenances must also be in the lexicon since, otherwise, prefix separation would take place and décontenances would be marked as a noun; listing only tenance would give contenance(s) the wrong features.)

avance	nnvb	forlance	v,sg
balance	nnvb	garance	nnvb
décontenance	v,sg	lance	nnvb
décontenances	v,sg	manigance	nnvb
devance	v,sg	mordance	v,sg
distance	nnvb	nuance	nnvb
élance	v,sg	ordonnance	nnvb
elance	v,sg	quittance	nnvb
fiance	v,sg	tance	v,sg
finance	nnvb		

60-Ending sme unchanged and marked noun, masculine, singular.

This word is not solely a noun:

enthousiasme	nnvb

61-Ending um unchanged and marked noun, masculine, singular.

The following word is not a masculine noun:

bégum	nn,f,sg

In this word, final um is not pronounced /ɔm/:

parfum	nn,m,sg

The term dum-dum is on the list of idioms.

62-Ending age unchanged and marked noun, singular.

The following words are singular verbs (dégages, déménages, emménages are listed since, otherwise, prefix separation would take place and they would be assigned the same features as gage and ménage; décourages, dédommages, dévisages are listed so that they will not be marked as nouns):

aménage	dégages	emménages	envisage
décourage	déménage	encage	propage
décourages	déménages	encourage	rengage
dédommage	dévisage	endommage	soulage
dédommages	dévisages	engage	surnage
dégage	emménage	enrage	

The following words belong to the category of noun/verb:

arrérage	fourrage	nage	partage
avantage	gage	naufrage	présage
étage	grillage	ombrage	rage
etage	ménage	outrage	ramage

ravage saccage voyage

The following word is an adverb:

davantage

63-Ending au unchanged and marked noun, masculine, singular.

The following word is not a masculine noun:

peau nn,f,sg

The words au and eau are on the list of idioms.

64-Ending aire unchanged and marked adjective, ambiguous gender, sing.

The following words are infinitives:

abstraire	extraire	plaire	surfaire
braire	faire	raire	taire
complaire	forfaire	rentraire	traire
contrefaire	malfaire	satisfaire	
distraire	parfaire	soustraire	

Other words which are not adjectives:

affaire	nnvb	flaire	v,sg
éclaire	v,sg	glaire	nnvb
eclaire	v,sg		

65-Ending ère unchanged and marked noun, singular.

The following words are singular verbs (blatère, carcère, cinère, fère, etc., yield correct features for déblatère(s), incarcère(s), incinère(s), con/dé/in/inter/pré/trans/fère(s), etc.; décolères is listed to prevent prefix separation which would result in décolères being marked as a noun):

accélère	digère	libère	réincarcère
acère	dilacère	lisère	réitère
acière	enquière	macère	rénumère
acquière	énumère	modère	révère
adhère	enumère	obère	sère
aère	espère	oblitère	sidère
agglomère	exagère	obtempère	suggère
altère	exaspère	opère	tempère
avère	exonère	persévère	tolère
blatère	fédère	pondère	vénère
carcère	fère	profère	vétère
cinère	génère	quière	vitupère
coopère	gère	récupère	vocifère
décolère	glomère	réfère	
décolères	lacère	réfrigère	
diffère	légifère	régénère	

Other words which are not adjectives:

adultère	adjnnvb	derrières	nn,m,pl
arrière	adjnadvb	prospère	adjnnvb
arrières	nnvb	repère	nnvb
congénère	adjnn,ag,sg	repères	nnvb
congénères	adjnn,ag,pl	réverbère	nnvb
congère	nn,f,sg	stère	nnvb
congères	nn,f,pl	ulcère	nnvb
derrière	nnprep;97,102,112		

66-Ending euse unchanged and marked noun, feminine, singular.

This word belongs to the category of adjective/verb: creuse.

67-Ending ble unchanged and marked adjective, ambiguous gender, sing.

The following words are singular verbs (dédoubles, démeubles, dessables, redoubles, remeubles are listed since, otherwise, prefix separation would take place and they would be assigned the same features as double, meuble, and sable; ressembles is listed so that it will not be marked adjective):

accable	dédoubles	endiable	redoubles	semble
affuble	démeuble	ensable	remeuble	
assemble	démeubles	entable	remeubles	
attable	dessable	rassemble	ressemble	
dédouble	dessables	redouble	ressembles	

Other words which are not adjectives:

amble	nnvb	etable	nnvb
câble	nnvb	jable	nnvb
comble	nnvb	meuble	adjnnvb
crible	nnvb	sable	nnvb
double	adjnnvb	table	nnvb
ensemble	nnadv	tremble	nnvb
ensembles	nn,m,pl	trouble	nnvb
étable	nnvb		

68-Ending ard changed to ar and marked noun, masculine, singular.

This word is an adverb: tard.

70-Ending ot changed to o and marked noun, masculine, singular.

The word Lot is on the list of ambiguous spellings.

71-Ending ant changed to an and marked adj., noun, or present participle.

Words which are not solely adjectives, nouns, or present participles:

durant	prepant	devants	nn,m,pl
devant	nprepant;97,102,112		

The words autant, avant, quant, tant are on the list of idioms.

72-Ending ins unchanged and marked noun, singular.

The following words are singular verbs:
(The words ending with ins must be in the lexicon since, otherwise, the table of endings would assign wrong linguistic features after the removal of final s.)

adjoins	empreins	eteins	peins
astreins	enceins	étreins	plains
atteins	enfreins	etreins	poins
ceins	enjoins	feins	ratteins
contrains	épreins	geins	restreins
crains	epreins	joins	teins
disjoins	éteins	oins	

Other words which are not nouns:

certain	adjnn,m,sg;94,95,106,108	plein	adj,m,sg
certains	adndp,m,pl,z;94,95,96,104	prochain	adj,m,sg
demain	adv	soudain	adv
divin	adj,m,sg	vain	adj,m,sg
enfin	adv	vilain	adj,m,sg
hein	adv		

The words afin and loin are on the list of idioms.

73-Ending iste unchanged and marked noun, singular.

The following words are singular verbs (contristes, dépistes, and désistes are included in order to prevent prefix separation; if it were allowed, contristes and dépistes would receive the same features as triste and piste, and désistes would be marked as a noun):

assiste	contristes	désistes	siste
attriste	dépiste	existe	subsiste
coexiste	dépistes	persiste	
contriste	désiste	résiste	

This word is a noun/verb: piste.

75-Ending oire unchanged and marked noun, singular.

Words which are not marked nouns:

accroire	inf	croire	inf
boire	nn,m,sg,npref,nninf	moire	nnvb

76-Ending ou unchanged and marked noun, masculine, singular.

Words which are not solely masculine singular nouns:
(The words ending with ous must be in the lexicon since, otherwise,

the table of endings would place them in the wrong category or mark them plural instead of ambiguous number after the removal of final s.)

absous	v,sg	entrevous	nn,m,an
bous	v,sg	nounou	adjnn,f,sg
burnous	nn,m,an	prou	adv
couscous	nn,m,an	remous	nn,m,an
dessous	nnadv	résous	v,sg
dissous	v,sg	sous	nnprep,z;97,102,112

The word tous is on the list of ambiguous spellings. The words nous and vous are on the list of idioms.

77-Ending an unchanged and marked noun, masculine, singular.

Words which are not solely masculine nouns:
(The words ending with ans must be in the lexicon since, otherwise, the table of endings would assign wrong features after final s is removed.)

céans	adv	dedans	nnadv
dans	prep,z;97,102,112	maman	nn,f,sg

In these masc. nouns, final an is pronounced /an/: jerrican, jerrycan.

The words Jean and sans are on the list of idioms.

81-Ending ès changed to èxx and marked noun, masc., ambiguous number.

Words which are not nouns:

profès	adj,m,an	très	adv,z

Masc. nouns with ambiguous number in which final s is not pronounced:

abcès accès agrès cyprès décès excès grès procès progrès succès

The words après, auprès, dès, lès, près are on the list of idioms. The word exprès is on the list of ambiguous spellings.

87-Remaining ending que changed to ke and marked adjective, noun, or verb, singular.

Words which are not adjectives, nouns, or verbs:
(The word oncques must be in the lexicon since, otherwise, the table of endings would assign wrong linguistic features after final s is removed.)

chaque	det,ag,sg;103	puisque	sub
jusque	prep	que	sub;110,111
lorsque	sub	quelconque	det,ag,sg
oncques	adv	quoique	sub
presque	adv		

The words quelque and quelques are on the list of idioms.

88-Remaining ending ue changed to u and marked adjective, noun, verb, or past participle, singular.

In this word, the ending ue must not be changed to u:

barbecue nn,m,sg

89-Ending nc changed to n and marked adjective, noun, or verb, sing.

In the following words, the ending nc must not be changed to n:

clinc	nn,m,sg	zinc	nn,m,sg

The word donc is on the list of ambiguous spellings.

90-Ending ing changed to ign and marked noun, masculine, singular.

In these masc. sing. nouns, the ending ing must not be changed to ign:

bastaing	coing	parpaing	seing	shampooing
basting	oing	poing	shampoing	shirting

The words contreseing and contreseings are in Group Three.

91-Remaining ending ng changed to n and marked noun, masc., singular.

In these masc. singular nouns, the ending ng must not be changed to n:

bang	gang	minnesang	rotang	slang
boomerang	ginseng	mustang	shantoung	
chantung	gong	pong	shantung	

93-Ending elle unchanged and marked adjective, noun, or verb, singular.
(The forms with a final s must be placed in the lexicon when their sets of linguistic features are not exactly the same as those of the forms without s.)

Words which are not adjectives, nouns, or verbs:

auxquelles	sub,f,pl;109,111	quelles	det,f,pl,z
desquelles	sub,f,pl;111	telle	det,f,sg
elles	allpron,f,pl,z;77,96,98,104	telles	det,f,pl,z
quelle	det,f,sg;108		

The words elle, celle and celles are on the list of idioms.

GROUP TWO

General comments:

1. This second group (1223 items) contains words (mostly high-frequency words) which facilitate the linguistic analysis of the input string.

Prefixed forms (see paragraph j, page 27) must be included whenever they do not belong to the same grammar category as the unprefixed words (e.g., délivre/s because of livre). The forms with a final s must be included (see section 5, page 85) when they do not belong to the same grammar category as the forms without s (e.g., être, êtres, tu, tus).

2. Words which appear in other groups (e.g., tiens in Group One, gens in Group Five, bien on the idi[illegible] list) are not included.

3. First, we list the words which belong to only one grammar category:

a. function words b. verbal forms c. nouns d. adjectives

and, then, we list the words which belong to several categories.

A. Function words

aujourdui	adv
auparavant	adv
auquel	sub,m,sg;109,111
aussi	adv
autrefois	adv
auxquels	sub,m,pl;109,111
avec	prep;97,102,112
bientôt	adv
çà	adv
cependant	adv
ces	det,ag,pl,z
cet	det,m,sg;103
cette	det,f,sg;103
ci	adv
çiy	spr;96
debout	adv
déjà	adv
desquels	sub,m,pl;111
dorénavant	adv
douze	nb
duquel	sub,m,sg;111
encore	adv
ensuite	coord
environ	adv
gratis	adv
guère	adv
huit	nb,tn
ici	adv
ils	spr,pl,z;104
jadis	adv
jiy	spr,sg
kiy	sub;110,111
longtemps	adv
lorsk	sub
me	di,ag,sg;84,88,90,100
mes	det,ag,pl,z
naguère	adv
néanmoins	adv
ni	coord;107
nos	det,ag,pl,z
notre	det,ag,sg;103
onze	nb
ou	coord
où	sub;109,111
parfois	adv
parmi	prep;97,112
partout	adv
pourtant	coord
puisk	sub
quatorze	nb
quel	det,m,sg;108
quelquefois	adv
quels	det,m,pl,z
qui	sub;104,111
quinze	nb
quoi	obpr;109
quoik	sub
sa	det,f,sg
se	di,ag,an;84,88,90,100
seize	nb
sept	nb
ses	det,ag,pl,z
sil	spr,sg
sils	spr,pl,z
soi	disjpr,ag,sg;98
sur	prep;97,102,112
ta	det,f,sg
tantôt	adv
tel	det,m,sg
tels	det,m,pl,z
tes	det,ag,pl,z
toujours	adv

toutefois	adv	une	det,f,sg;103,106	voilà	prep;112
treize	nb	vite	adv	vos	det,ag,pl,z
trois	nb,z	voici	prep;112	votre	det,ag,sg;103

B. Verbal forms

All the verbs ending with:

age	ages	ai	aire	aires	ais	ait
ance	ances	ante	antes	ble	bles	end
ends	ent	er	ère	ères	ers	eurs
euse	euses	ez	ienne	iennes	iens	ins
ionne	ionnes	ique	iques	ir	iste	istes
ndre	ndres	oir	oire	oires	ont	ous
ra	ras	rut	sme	smes	tions	ux

have already been marked as verbs by the table of endings and, therefore, are not included in the following list (except for veulent which must be included so that veule and veulent will be pronounced correctly).

The table of endings and the lexicon find all the French words which are solely infinitives (including rare verbs such as chaloir and sourdre) and the few words which are noun/infinitive (e.g., rire, aller) or adjective/infinitive (e.g., tendre).

abattre	inf	avons	v,sg	concasses	v,sg
absoudre	inf	ayant	prespar	conclure	inf
accroître	inf	ayons	v,sg	conduire	inf
achète	v,sg	bat	v,sg	confire	inf
acheté	ptp	battre	inf	connaisse	v,sg
admettre	inf	battu	ptp,m,sg	connaît	v,sg
aille	v,sg	boit	v,sg	connaître	inf
aime	v,sg	bruire	inf	connu	ptp
aimé	ptp	bu	ptp	connut	v,sg
aimons	v,pl	buvant	prespar	consomme	v,sg
alla	v,sg	cherchant	prespar	consommes	v,sg
allé	ptp,m,sg	cherche	v,sg	construire	inf
allons	v,pl	cherché	ptp	contredire	inf
amène	v,sg	choisi	ptp	corrompre	inf
amené	ptp	choisisse	v,sg	coudre	inf
aperçoit	v,sg	choisit	v,sg	courre	inf
aperçut	v,sg	circoncire	inf	couru	ptp
apparaître	inf	clore	inf	coûtant	prespar
appartenant	prespar	clut	v,sg	coûte	v,sg
appartenu	ptp	combattre	inf	coûté	ptp
appelle	v,sg	commence	v,sg	couvrant	prespar
arrive	v,sg	commencé	ptp	couvre	v,sg
arrivé	ptp,m,sg	commettre	inf	croit	v,sg
assied	v,sg	comparaître	inf	croître	inf
assoit	v,sg	compris	ptp	cuire	inf
attendant	prespar	compromettre	inf	déboîte	v,sg
attende	v,sg	compté	ptp	déboîtes	v,sg
aurons	v,sg	concasse	v,sg	débouche	v,sg

débouches	v,sg	donnant	prespar	êtes	v,pl
déduire	inf	donné	ptp,m,sg	etes	v,pl
défaille	v,sg	dû	ptp	eut	v,sg
défailles	v,sg	duré	ptp,m,sg	eût	v,sg
défeuille	v,sg	dut	v,sg	exclure	inf
défeuilles	v,sg	ébattre	inf	faisions	v,pl
dégorge	v,sg	ebattre	inf	faisons	v,pl
dégorges	v,sg	éclore	inf	faites	v,pl
dégrise	v,sg	eclore	inf	fallu	ptp
dégrises	v,sg	éconduire	inf	fallut	v,sg
déjeune	v,sg	econduire	inf	fasse	v,sg
déjeuné	ptp	écoutant	prespar	fende	v,sg
déjeunes	v,sg	ecoutant	prespar	fendu	ptp
déjoue	v,sg	écouté	ptp	fermant	prespar
déjoues	v,sg	ecouté	ptp	fermé	ptp
délaisse	v,sg	écrire	inf	finissant	prespar
délaisses	v,sg	ecrire	inf	finisse	v,sg
délivre	v,sg	écrivant	prespar	finit	v,sg
délivres	v,sg	ecrivant	prespar	fit	v,sg
demandé	ptp	écrive	v,sg	forclore	inf
démets	v,sg	ecrive	v,sg	frire	inf
dénombre	v,sg	élire	inf	fume	v,sg
dénombres	v,sg	elire	inf	fumé	ptp,m,sg
dénote	v,sg	emboîte	v,sg	fut	v,sg
dénotes	v,sg	emboîtes	v,sg	gagne	v,sg
dépasse	v,sg	émettre	inf	gagné	ptp
dépasses	v,sg	emettre	inf	gardant	prespar
dépensé	ptp,m,sg	émoudre	inf	gardé	ptp
déplace	v,sg	emoudre	inf	gèle	v,sg
déplaces	v,sg	empêche	v,sg	habillant	prespar
déporte	v,sg	empêché	ptp	habille	v,sg
déportes	v,sg	empoche	v,sg	habite	v,sg
dépose	v,sg	empoches	v,sg	habité	ptp
déposes	v,sg	emporte	v,sg	habitue	v,sg
descende	v,sg	emportes	v,sg	haïr	inf
descendu	ptp	enclore	inf	inclure	inf
désherbe	v,sg	enduire	inf	induire	inf
désherbes	v,sg	ensuivre	inf	instruire	inf
détenant	prespar	entendu	ptp	intéresse	v,sg
détournant	prespar	entrant	prespar	interpose	v,sg
détrompe	v,sg	entré	ptp,m,sg	interposes	v,sg
détrompes	v,sg	entrepose	v,sg	interrompre	inf
détruire	inf	entreposes	v,sg	introduire	inf
devenu	ptp	entres	v,sg	invite	v,sg
devons	v,sg	entretenant	prespar	jetant	prespar
dis	v,sg	entrevit	v,sg	jeté	ptp,m,sg
disant	prespar	envoie	v,sg	jette	v,sg
dise	v,sg	es	v,sg	jouant	prespar
disparaître	inf	essaie	v,sg	joué	ptp
dissoudre	inf	essayant	prespar	laissant	prespar
dites	v,pl	essayé	ptp	laissé	ptp
doit	v,sg	étant	prespar	lave	v,sg
doive	v,sg	etant	prespar	lavé	ptp

levé	ptp,m,sg	parlé	ptp	reçoit	v,sg
lève	v,sg	parlons	v,pl	reçoive	v,sg
lire	inf	parte	v,sg	reçut	v,sg
lisant	prespar	paru	ptp	redemande	v,sg
lise	v,sg	passée	ptp,f,sg	redemandes	v,sg
lu	ptp	payant	prespar	redonne	v,sg
luire	inf	payé	ptp	redonnes	v,sg
mange	v,sg	pensant	prespar	réduire	inf
mangé	ptp	pense	v,sg	réélire	inf
mangeant	prespar	pensé	ptp,m,sg	regardant	prespar
manqué	ptp	perd	v,sg	regarde	v,sg
marchant	prespar	perdre	inf	regardé	ptp
maudire	inf	perdu	ptp	regorge	v,sg
méconnaître	inf	permet	v,sg	regorges	v,sg
médire	inf	permettre	inf	regrette	v,sg
mentant	prespar	pesé	ptp,m,sg	regretté	ptp
mente	v,sg	pèse	v,sg	rejoue	v,sg
menti	ptp	peut	v,sg	rejoues	v,sg
met	v,sg	plaît	v,sg	relit	v,sg
mettant	prespar	pleurant	prespar	remarche	v,sg
mette	v,sg	pleure	v,sg	remarches	v,sg
mettre	inf	pleuré	ptp	remets	v,sg
meure	v,sg	pleut	v,sg	rempli	ptp
meurt	v,sg	pleuve	v,sg	remplisse	v,sg
mis	ptp	plu	ptp	remplit	v,sg
monté	ptp,m,sg	plut	v,sg	rencontrant	prespar
moque	v,sg	porté	ptp,m,sg	rencontré	ptp
moqué	ptp	posé	ptp	rende	v,sg
moudre	inf	poursuivre	inf	renonce	v,sg
naître	inf	pouvant	prespar	renoncé	ptp
naquit	v,sg	préféré	ptp	repaie	v,sg
né	ptp	prenant	prespar	repaies	v,sg
neigé	ptp	présenté	ptp	repart	v,sg
nuire	inf	prêtant	prespar	repasse	v,sg
obéi	ptp	prêté	ptp	repasses	v,sg
obéisse	v,sg	prévit	v,sg	replace	v,sg
obéit	v,sg	produire	inf	replaces	v,sg
occire	inf	promet	v,sg	répond	v,sg
occlure	inf	promettre	inf	réponde	v,sg
occupe	v,sg	pu	ptp	répondu	ptp
occupé	ptp	puisse	v,sg	reporte	v,sg
omettre	inf	put	v,sg	reportes	v,sg
ont	v,pl,t	quittant	prespar	repose	v,sg
ose	v,sg	quitté	ptp	reposé	ptp
oublié	ptp	rabattre	inf	reposes	v,sg
ouïr	inf	réadmettre	inf	résoudre	inf
ouvert	ptp	réapparaître	inf	ressemblé	ptp
ouvrant	prespar	rebouche	v,sg	resté	ptp
ouvre	v,sg	rebouches	v,sg	retombe	v,sg
paître	inf	recasse	v,sg	retombes	v,sg
paraît	v,sg	recasses	v,sg	retournant	prespar
paraître	inf	recevant	prespar	retourne	v,sg
parle	v,sg	reçois	v,sg	retourné	ptp

retranche	v,sg	sourdre	inf	traversant	prespar
retranches	v,sg	souri	ptp	traversé	ptp,m,sg
retraverse	v,sg	souriant	prespar	trompé	ptp
retraverses	v,sg	sourit	v,sg	trouve	v,sg
réuni	ptp	souvenu	ptp	trouvé	ptp
réunisse	v,sg	soyons	v,pl	tut	v,sg
réunit	v,sg	su	ptp	va	v,sg
réussi	ptp	suffire	inf	vaille	v,sg
réussisse	v,sg	suis	v,sg,z	vaincre	inf
réussit	v,sg	suit	v,sg	valant	prespar
revit	v,sg	suive	v,sg	valu	ptp,m,sg
revive	v,sg	suivi	ptp	valut	v,sg
revives	v,sg	suivre	inf	vaut	v,sg
ri	ptp	survivre	inf	vécu	ptp
rie	v,sg	sut	v,sg	vécut	v,sg
rompre	inf	taisant	prespar	venant	prespar
sachant	prespar	taise	v,sg	vendant	prespar
sache	v,sg	téléphoné	ptp	vende	v,sg
sali	ptp	tenu	ptp,m,sg	vendu	ptp
salit	v,sg	tombé	ptp,m,sg	venu	ptp,m,sg
séduire	inf	tourne	v,sg	veuille	v,sg
sentant	prespar	tourné	ptp,m,sg	veulent	v,pl
senti	ptp	traduire	inf	veut	v,sg
serons	v,pl	traduisant	prespar	vidant	prespar
sert	v,sg	traduise	v,sg	vidé	ptp
serve	v,sg	transmets	v,sg	visité	ptp
servi	ptp	transporte	v,sg	voit	v,sg
sois	v,sg	transportes	v,sg	vole	v,sg
sont	v,pl,t	transpose	v,sg	voté	ptp
sors	v,sg	transposes	v,sg	voulu	ptp
sorti	ptp,m,sg	transvide	v,sg	voulut	v,sg
souffert	ptp	transvides	v,sg	voyageant	prespar
souffrant	prespar	travaillant	prespar	voyagé	ptp
souffre	v,sg	travaille	v,sg	vu	ptp,m,sg
soumettre	inf	travaillé	ptp		

C. Nouns

Nouns which end with the following strings:

ents	ais	ont	eur	um	ot	erie
oiement	ai	onts	eurs	ums	ots	eries
alement	ez	end	ic	age	in	ogie
ellement	ix	ends	ics	ages	ins	ogies
eusement	tions	ra	ien	au	iste	ès
gement	on	ras	iens	aus	istes	ss
onnement	ers	ait	ienne	ère	oire	cratie
ssement	er	aits	iennes	ères	oires	craties
chement	irs	rut	ance	euse	ou	ng
ivement	ir	ruts	ances	euses	ous	ngs
all remaining ent	ndre	ionne	sme	ard	an	
except ement	ndres	ionnes	smes	ards	ans	

have already been marked as such by the table of endings and, therefore, are not included in the list below. Also, we are not including the nouns which appear in other groups (e.g., nerf, outil, sandwich which are in Group Three; fils, gens which are in Group Five; corps, mot, nuit, père which are part of the idiom list) or the nouns which are composed of a prefix plus a noun on the list (e.g., prénom).

accord	m,sg
aérogare	f,sg
aérogares	f,pl
âge	m,sg
age	m,sg
allemand	m,sg
allemande	f,sg
allumette	f,sg
ami	m,sg
amie	f,sg
amour	m,sg
animal	m,sg
animaux	m,pl
année	f,sg
antichambre	f,sg
antichambres	f,pl
août	m,sg
appartement	m,sg
arbre	m,sg
avants	m,pl
avril	m,sg
bagages	m,pl
banque	f,sg
bateaux	m,pl
bibliothèque	f,sg
bicyclette	f,sg
bifteck	m,sg
bijoux	m,pl
billet	m,sg
boîte	f,sg
bouteille	f,sg
bouts	m,pl
bruit	m,sg
ça	sg;96,106
café	m,sg
cailloux	m,pl
campagne	f,sg
capitale	f,sg
ceci	sg;106
cela	sg;106
centime	m,sg
centimètre	m,sg
chacun	m,sg
chacune	f,sg
chaise	f,sg
chapeaux	m,pl
chaussette	f,sg
chaussure	f,sg
cheval	m,sg
chevaux	m,pl
cheveu	m,sg
cheveux	m,pl
chocolat	m,sg
chose	f,sg
choux	m,pl
ciel	m,sg
cieux	m,pl
cigarette	f,sg
cinéma	m,sg
clé	f,sg
clef	f,sg
confort	m,sg
conforts	m,pl
conque	f,sg
conques	f,pl
conseil	m,sg
consort	m,sg
convive	ag,sg
convives	ag,pl
côté	m,sg
coup	m,sg
courrier	m,sg
courriers	m,pl
couteaux	m,pl
crédit	m,sg
crème	f,sg
décembre	m,sg
délit	m,sg
dessert	m,sg
dévote	f,sg
dévotes	f,pl
dimanche	m,sg
dires	m,pl
doigt	m,sg
dollar	m,sg
douzaine	f,sg
école	f,sg
ecole	f,sg
église	f,sg
eglise	f,sg
emprise	f,sg
emprises	f,pl
enfant	ag,sg
entremets	m,pl
espagnol	m,sg
espagnole	f,sg
état	m,sg
etat	m,sg
êtres	m,pl
etres	m,pl
examen	m,sg
famille	f,sg
femme	f,sg
fenêtre	f,sg
feu	m,sg
feuille	f,sg
feux	m,pl
fille	f,sg
film	m,sg
fleuve	m,sg
fois	f,an
fourchette	f,sg
foutres	m,pl
française	f,sg
fruit	m,sg
fûts	m,pl
gâteaux	m,pl
genoux	m,pl
géographie	f,sg
gramme	m,sg
guerre	f,sg
habitude	f,sg
herbe	f,sg
heure	f,sg
homme	m,sg
hôpital	m,sg
hôpitaux	m,pl
hôtel	m,sg
idée	f,sg
jambe	f,sg
jeu	m,sg
jeudi	m,sg
joie	f,sg
jour	m,sg
journal	m,sg
journaux	m,pl
journée	f,sg
juillet	m,sg

jus	m,an	nuits	f,pl	savant	m,sg
kilo	m,sg	numéro	m,sg	sel	m,sg
kilogramme	m,sg	octobre	m,sg	semaine	f,sg
kilomètre	m,sg	oeil	m,sg	septembre	m,sg
langue	f,sg	oeuf	m,sg	serviette	f,sg
laquelle	f,sg;106	oeufs	m,pl	siècle	m,sg
légume	ag,sg	oiseaux	m,pl	ski	m,sg
lequel	m,sg;106	oncle	m,sg	soie	f,sg
lesquelles	f,pl;104	orange	f,sg	soif	f,sg
lesquels	m,pl;104	oreille	f,sg	soirée	f,sg
lettre	f,sg	parts	f,pl	sol	m,sg
lieu	m,sg	pays	m,an	soleil	m,sg
lieux	m,pl	pendants	m,pl	sons	m,pl
litre	m,sg	personnes	f,pl	sorts	m,pl
lits	m,pl	photo	f,sg	sourires	m,pl
loi	f,sg	pièce	f,sg	stylo	m,sg
lundi	m,sg	pied	m,sg	surnom	m,sg
lunette	f,sg	pluie	f,sg	tabac	m,sg
maisons	f,pl	pneu	m,sg	tarte	f,sg
manteaux	m,pl	pois	m,an	taxi	m,sg
mardi	m,sg	pomme	f,sg	temps	m,an
mari	m,sg	poulet	m,sg	tennis	m,an
mars	m,an	printemps	m,an	tête	f,sg
matinée	f,sg	pronom	m,sg	thé	m,sg
maux	m,pl	quadrique	f,sg	théâtre	m,sg
mensonge	m,sg	quadriques	f,pl	tons	m,pl
mercredi	m,sg	quart	m,sg	touts	m,pl
mètre	m,sg	quelkun	m,sg;106	travail	m,sg
métro	m,sg	quelkune	f,sg;106	travaux	m,pl
midi	m,sg	quiconque	ag,sg	vacances	f,pl
milieu	m,sg	radio	f,sg	valise	f,sg
millions	m,pl	repas	m,an	vélo	m,sg
minuit	m,sg	réponse	f,sg	vendredi	m,sg
mois	m,an	restaurant	m,sg	vérité	f,sg
moitié	f,sg	retard	m,sg	verre	m,sg
monde	m,sg	retards	m,pl	veste	f,sg
montagne	f,sg	rires	m,pl	vêtement	m,sg
mur	m,sg	riz	m,an	viande	f,sg
neveu	m,sg	route	f,sg	vie	f,sg
neveux	m,pl	russe	ag,sg	ville	f,sg
nièce	f,sg	salade	f,sg	vinaigre	m,sg
nom	m,sg	salle	f,sg	vits	m,pl
nourriture	f,sg	samedi	m,sg	vivres	m,pl
novembre	m,sg	santé	f,sg	yeux	m,pl

D. Adjectives

Adjectives which end with the following strings:

ents	ons	ndre	rut	ien	vienne	aire	al	ièmes
ent	on	ndres	ruts	iens	viennes	aires	in	if
ais	ers	ra	pique	tienne	ante	ble	ins	ifs
ai	er	ras	piques	tiennes	antes	bles	ième	ès

have already been marked as such by the table of endings and, therefore, are not included in the list below. Also, we are not including the adjectives which appear in other groups (e.g., abrupt, adéquat, distinct which are in Group Three; content, bis which are in Group Five; sauf which is part of the idiom list) or the adjectives which are composed of a prefix plus an adjective on the list (e.g., déshonnête).

affreux	m,an	gentille	f,sg	nouvel	m,sg
âgé	sg	grand	m,sg	nouvelle	f,sg
agé	sg	grande	f,sg	pauvre	ag,sg
américaine	f,sg	grasse	f,sg	petite	f,sg
ancienne	f,sg	gris	m,an	pire	ag,sg
anglaise	f,sg	gros	m,an	pleine	f,sg
antineige	ag,sg	grosse	f,sg	principaux	m,pl
antineiges	ag,pl	haute	f,sg	prochaine	f,sg
assise	f,sg	honnête	ag,sg	proches	ag,pl
antique	ag,sg	inaperçu	m,sg	profond	m,sg
antiques	ag,pl	inaperçue	f,sg	profonde	f,sg
basse	f,sg	inaperçues	f,pl	propre	ag,sg
beaux	m,pl	inaperçus	m,pl	quadriparti	m,sg
bel	m,sg	infini	sg	quadripartie	f,sg
blanche	f,sg	infinis	pl	quadriparties	f,pl
bleu	sg	jaune	ag,sg	quadripartis	m,pl
célèbre	ag,sg	jeune	ag,sg	quiet	m,sg
chaud	m,sg	joli	sg	quiète	f,sg
chaude	f,sg	large	ag,sg	riche	ag,sg
courte	f,sg	lente	f,sg	rouge	ag,sg
demi	sg	long	m,sg	sale	ag,sg
difficile	ag,sg	longue	f,sg	saufs	m,pl
divine	f,sg	lourde	f,sg	sec	m,sg
doux	m,an	maigre	ag,sg	sèche	f,sg
droite	f,sg	malade	ag,sg	seul	m,sg
dur	m,sg	mauvaise	f,sg	seule	f,sg
épaisse	f,sg	meilleur	m,sg	sis	m,an
epaisse	f,sg	meilleure	f,sg	suivants	m,pl
facile	ag,sg	mêmes	ag,pl	tranquille	ag,sg
feue	f,sg	mince	ag,sg	unique	ag,sg
fol	m,sg	moyen	m,sg	uniques	ag,pl
folle	f,sg	moyenne	f,sg	utile	ag,sg
fortes	f,pl	multicarte	ag,sg	vaine	f,sg
forts	m,pl	multicartes	ag,pl	vert	m,sg
fraîche	f,sg	multiplace	ag,sg	verte	f,sg
froid	m,sg	multiplaces	ag,pl	vieil	m,sg
froide	f,sg	neuve	f,sg	vieille	f,sg
gauche	ag,sg	nombreux	m,pl	vieux	m,an
générale	f,sg	nouveau	m,sg	vilaine	f,sg
gentil	m,sg	nouveaux	m,pl	vraie	f,sg

E. Words which belong to several grammar categories

The words in this group (mostly high-frequency words) belong to several grammar categories (e.g., place which is a verb or a noun).

Some of the words in this group had to be listed (e.g., empire) because they are prefixed words which do not have the same grammar categories as those of the unprefixed words (e.g., pire). Others (e.g., tombée) had to be included because their masculine form is listed in the lexicon as belonging to only one grammar category.

Placing a word on this list is of value only if one of these conditions is satisfied:

- Our list assigns linguistic features which are more precise than those assigned by the table of endings; for example, rue is marked "nnvb" in the list below instead of being marked "adjnnvb" by the table of endings.

- The table of endings would not assign any linguistic features (e.g., beurre, court, chemise).

1. Words which are given more precise linguistic features than those assigned by the table of endings:

allée	nnpastpr,f,sg	partant	nnant
arrivée	nnpastpr,f,sg	partants	adjnn,m,pl
battue	nnpastpr,f,sg	pensée	nnpastpr,f,sg
belle	adjnn,f,sg	pesée	nnpastpr,f,sg
dépassée	adjpp,f,sg	portée	adjnnpp,f,sg
dépassées	adjpp,f,pl	rejetée	adjpp,f,sg
dépensée	adjpp,f,sg	rejetées	adjpp,f,pl
dépensées	adjpp,f,pl	remariée	adjpp,f,sg
dépolie	adjpp,f,sg	remariées	adjpp,f,pl
dépolies	adjpp,f,pl	repesée	adjpp,f,sg
donnée	nnpastpr,f,sg	repesées	adjpp,f,pl
durée	nnpastpr,f,sg	répondant	nnant
empesée	adjpp,f,sg	répondants	adjnn,m,pl
empesées	adjpp,f,pl	reportée	adjpp,f,sg
entrée	nnpastpr,f,sg	reportées	adjpp,f,pl
finie	adjpp,f,sg	rue	nnvb
fumée	nnpastpr,f,sg	sortie	nnpastpr,f,sg
gelée	adjnnpp,f,sg	sue	vbpastpr
habitant	nnant	tenant	nnant
habitants	nn,m,pl	tenants	nn,m,pl
huile	nnvb	tenue	nnpastpr,f,sg
jetée	nnpastpr,f,sg	tombée	nnpastpr,f,sg
joue	nnvb	tournant	adjnnant
laine	nnvb	tournants	adjnn,m,pl
levée	nnpastpr,f,sg	tournée	nnpastpr,f,sg
manque	nnvb	transportée	adjpp,f,sg
marque	nnvb	transportées	adjpp,f,pl
monnaie	nnvb	traversée	nnpastpr,f,sg
montée	nnpastpr,f,sg	value	nnpastpr,f,sg
mourant	nnant	venue	nnpastpr,f,sg
mourants	adjnn,m,pl	volée	nnpastpr,f,sg
paie	nnvb	vue	nnpastpr,f,sg

2. Words which would not be found by the table of endings:

a	nnvb
adresse	nnvb
aperçu	nnpastpr,sg
appelé	nnpastpr,sg
assis	adjnnpp,m,an
attendu	nnpastpr,sg
aucun	dp,m,sg,n;94,95,96,103,106
aucune	dp,f,sg;94,95,96,103,106
aucunes	dp,f,pl,z;96,104
aucuns	dp,m,pl,z;96,104
autre	adndp,ag,sg
autres	adndp,ag,pl,z
aveugle	adjnnvb
bas	nnadjadv,z
beurre	nnvb
blond	adjnn,m,sg
blonde	adjnn,f,sg
bonne	adjnn,f,sg
bouche	nnvb
brun	adjnn,m,sg
brune	adjnn,f,sg
carte	nnvb
casse	nnvb
certaine	adjnn,f,sg;94,95,106,108
certaines	adndp,f,pl,z;94,95,96,104
chambre	nnvb
chemise	nnvb
contre	nnvbprep;97,102,112
contres	nnvb
cours	nnvb
court	adjnadvb
couvert	nnpastpr,m,sg
cravate	nnvb
cru	adjnnpp,sg
cuisine	nnvb
cuit	adjnnvb
cuite	nnpastpr,f,sg
cuits	adjpp,m,pl
date	nnvb
débat	nnvb
débats	nnvb
défini	adjpp,m,sg
définis	adjvb
dehors	nnadv
demande	nnvb
dépassé	adjnnpp,m,sg
dépassés	adjnnpp,m,pl
dépense	nnvb
dépenses	nnvb
dépoli	adjpp,m,sg
dépolis	adjvb
déshabillé	adjnnpp,m,sg
dessus	nnadv
dire	nninf
dit	vbpastpr
diverses	adndp,f,pl,z
donne	nnvb
dure	adjnnvb
écoute	nnvb
ecoute	nnvb
écrit	adjnvbpp
ecrit	adjnvbpp
écrits	adjnnpp,m,pl
ecrits	adjnnpp,m,pl
empire	nnvb
empires	nnvb
entre	vbprep;97
envoyé	nnpastpr,sg
être	nninf
etre	nninf
excepté	vbprep
faille	nnvb
fatigué	adjnnpp,sg
ferme	nnvb
fini	adjnnpp,sg
finis	adjvb
fort	nnadjadv,t
foutre	nninf
fraise	nnvb
fût	nnvb
garde	nnvb
gare	nnvb
gelé	adjpp,m,sg
glace	nnvb
gorge	nnvb
grise	adjvb
habillé	adjpp,sg
habitué	nnpastpr,sg
haut	nnadjadv,t
intéressé	nnpastpr,sg
invité	adjnnpp,sg
jalouse	adjnnvb
laisse	nnvb
lit	nnvb
lourd	nnadjadv
lourds	adjnn,m,pl
machine	nnvb
mal	nnadv
marche	nnvb
marché	nnpastpr,sg
marié	adjnnpp,sg
même	adjadv
mets	nnvb
minute	nnvb

moi	allpron,ag,sg;79,98
mort	adjnnpp,m,sg
morte	adjnnpp,f,sg
mouillé	adjnnpp,sg
neige	nnvb
nombre	nnvb
note	nnvb
nul	adndp,m,sg;94,95,96,106,108
nulle	adndp,f,sg;94,95,96,106,108
osé	adjnnpp,sg
oublie	nnvb
outre	nnvbprep;97,102
outres	nnvb
ouverte	adjnnpp,f,sg
parlant	adjnnant,m,sg
parlants	adjnn,m,pl
part	nnvb
parti	nnpastpr,sg
passe	nnvb
passé	adjnnpp,m,sg
paye	nnvb
peint	vbpastpr
permis	adjnvbpp
personne	nnadv;96,106
plaint	vbpastpr
poche	nnvb
point	nnvbadv,t
points	nnpastpr,m,pl
poli	adjnnpp,sg
polis	adjnvbpp
porte	nnvb
pose	nnvb
poste	nnvb
présente	nnvb
pressé	adjpp,sg
prête	adjvb
pris	adjnvbpp
prise	adjnnpp,f,sg
promis	adjnvbpp
pus	nnvb
recouvert	adjpp,m,sg
recouverts	adjpp,m,pl
reçu	nnpastpr,sg
remarié	adjpp,m,sg
remariés	adjpp,m,pl
rencontre	nnvb
rendu	nnpastpr,sg
reste	nnvb
rire	nninf
robe	nnvb
sente	nnvb
sommes	nnvb
sort	nnvb
sorte	nnvb
sourire	nninf
souris	nnvb
sucre	nnvb
tasse	nnvb
téléphone	nnvb
timbre	nnvb
toi	allpron,ag,sg;79,98
tombe	nnvb
traduit	vbpastpr
tranche	nnvb
traverse	nnvb
trompe	nnvb
tu	pronpp;77,96
tus	vbpastpr
usé	adjnnpp,sg
vide	adjnnvb
visite	nnvb
vit	nnvb
vive	adjvb
vivre	nninf
voisine	adjnnvb
voiture	nnvb
volé	adjnnpp,m,sg
vote	nnvb

GROUP THREE

This group contains 2032 words with special pronunciation problems. It does not contain all the words with such problems because some words appear in other groups (e.g., cuiller in Group One, desquels in Group Two, fils in Group Five, porc on the idiom list) or because their difficulties are solved by the root table (e.g., eusse) or by the transcription rules (e.g., baptême). Except for the transcription of a and â (see page 196), our program uses the same transcriptions as the Petit Robert; for the words which are not in the Petit Robert (e.g., nonos), we use other reliable dictionaries. The linguistic features are set as in Groups One and Two; the only words which list a feature above no. 57 are auxdites, auxdits, desdites, desdits, lesdits, lesdites which show feature 65 (/z/ of liaison).

abbaye
abc
abdomen
abribus
abrupt
acarus
accessit
accroc
achillée
acholie
achondroplasie
achromat
aconit
acrobatie
addenda
adéquassion
adéquat
adéquate
adonis
égosome
épyornis
éschne
affidavit
affûtiaux
agenda
agnus
agrostis
agrotis
ah
aïe
aiguillier
aiguilliers
airbus
albatros
albinos
albumen
alcarazas
alias
alios
allégrement
allegretto
allegrettos
allegro
allegros
alléluia
almanach
almasilium
altocumulus
altostratus
altuglas
amaryllis
amazone
âme
ame
amen
american
amict
amnios
amniosynthèse
amphioxus
ananas
anéchoïque
anet
aneth
angélus
anilinctus
anis
anthémis
antibois
anus
aoûtien
apartheid
appas
appât
appâte
aqua
aquicole
aquiculteur
aquicultrice
aquiculture
aquifère
archal
archégone
archéosocial
archéosociale
archéosociales
archéosocialisme
archéosociaux
archétype
archiatre
archiatres
archivent
archiver
archonte
arguë
arguënt
arguës
argus
argutie
arome
arroyo
arte
artefact
asialie
asparagus
aspect
aspergille
aspergillose
aspergillus
atlas
atome
âtre
atre
auburn
aulnaie
aulne
aulx
aurochs
auspice
autobus
automne
automnes
autours
auxdites
auxdits
auxerrois
auxerroise
auxerroises
auxois
auxoise
auxoises
ave
aven
avunculaire
axiome
axis
ayatollah
azimut
azygos
b
bâbord
bâbordaise
baby
bacul
bah
balletomane
barman
barmen
baroud
basbleu
bastos
bât
batayole
bâte
bath
bazooka
beagle
because
behaviorisme
behavioriste
belouga
ben
bene
bengali
benjamin
benjamine
benjoin
benthique
benthos
bentonite
béotien
biacide
biacides
biacuminé
biaural
biaurale
biauraux
biauriculaire
bibendum
bibliobus
biceps
bichlamar
bienheureuse
bienheureuses
bienheureux
biennal
biennale
biennales
biennaux
bigouden
billevesée
billion
bingo
binz
biotes
bioxyde
bioxydes
biscayen
biscayenne
bismuth
bit
bitmètre
bitos
biunivoque
blâme
blaps
bled
blende
bleuetière
blinis
blitzkrieg
blockhaus

blocus
blues
boat
bobo
bobsleigh
boësse
boëte
boette
boëtte
boeufs
bonneterie
bonshommes
bonus
booléen
booléenne
boolien
boolienne
boom
bougainvillée
bougainvillié
boutillié
bowling
boy
boyard
boyscout
brasero
break
breakfast
brechtien
brechtienne
bren
brichton
briqueterie
broc
brome
brushing
bruxellois
bruxelloise
bruyère
buco
building
bungalow
business
byssus
c
cabas
cacaoyers
cactus
cadenas
caennaise
cafeteria
caféteria
caïd
cake
calendos
caloyère
caloyers
calvados
calville
calvitie
cameraman
cameramen
campus
cancoillotte
canitie
cannabis
canoë
cantabile
cantaloup
caoutchouc
capésien
capésienne
capétien
capillarité
carborundum
carolus
castus
casus
casyllabe
casyllabique
catéchuménat
catéchumène
catgut
catharsis
catoblépas
celluloïd
cens
cept
cerf
cérumen
chalaze
chalazion
chalcographie
chalcopyrite
chalcosine
chaldéen
chaldéenne
chamérops
chamito
chaos
chaotique
charadriidé
chargeure
charismatique
charisme
cheiroptère
chéiroptère
chelem
chélicère
chélidoine
chéloïde
chélonien
chenal
chenaux
chénopode
chénopodiacée
chenu
chète
chevêche
chevesne
chevêtre
cheville
cheviotte
chianti
chiasma
chiasme
chiendents
chinchilla
chinook
chintz
chips
chitine
chitineuse
chitineux
chiton
chlamyde
chloasma
choatif
choative
choéphore
choeur
chondriome
chondriosome
chondroblaste
chorizo
chorizos
chorus
chrétienté
chromage
chrome
chromosome
chroumer
chthonien
chthonienne
ciao
cipaye
circonspect
cirrus
citrus
clancul
clausus
clebs
clerc
clitoris
clonus
clown
clownerie
clownesque
club
coaltar
cobaye
coccyx
cochylis
cockpit
cocktail
coco
codicille
coït
coitus
colibacille
colibacillose
collapsus
columbarium
compendium
compound
comput
conchylis
condom
condoms
condottiere
condottieri
confier
consensus
conter
contes
contreclef
contrecoup
contreseing
contreseings
contresens
convolvulus
cool
coolie
copyright
coquillié
coquillière
coréopsis
corpus
corticosurrénale
corticosurrénalien
corticosurrénalienne
cosaque

cosinus
cosmodrome
cosmos
cosy
couguar
coutelas
coutil
cow
coyau
coyote
crabe
craignos
craonaise
craps
cream
credo
crescendo
crésus
cricket
crocus
cromlech
crosne
crouchti
cubitus
cucul
cul
cumulus
cunnilinctus
cunnilingus
curaçao
cursus
cycas
cyclamen
cyclone
cynips
cyrillique
czar
czarine
d
dam
dan
deça
décanter
decauville
déchut
deci
déciment
décorner
decrescendo
défier
dei
delà
deleatur
delirium
démentions
dendrite
dendritique
dengue
dénotions
déparent
déplacer
déportions
derechef
desdites
desdits
désemparent
desiderata
desperado
détalent
détritus
deuxième
deuxièmement
deuxio
dévient
diable
diesel
digit
diktat
dinosaurien
diplomatie
disaccharide
discount
dissert
distinct
dixieland
dixième
dixièmement
dixit
djebel
dodécasyllabe
dodécasyllabique
dodo
dolce
dolichocéphale
dolmen
dom
doris
dot
doullennaise
doulos
drap
drink
drive
drome
duchnoc
duchnoque
dumping
duodenum
duramen
durit
dylome
e
échidné
echidné
échinocactus
echinocactus
échinococcose
echinococcose
échinocoque
echinocoque
échinoderme
echinoderme
écho
echo
échographie
echographie
échographique
echographique
écholalie
echolalie
écholocassion
echolocassion
échotié
echotié
échotière
echotière
edelweiss
éden
eden
effendi
efferalgan
ego
égyptien
egyptien
eh
eidétique
électrodes
electrodes
éléis
eléis
éléphantiasis
eléphantiasis
emmenthal
emmenthals
emmi
emparent
empâte
emportions
empressent
encas
encoignure
ennéa
ennéade
ennéades
ennemi
entrelacs
éphod
ephod
épispadias
epispadias
épistaxis
epistaxis
eptie
équanimité
equanimité
équassion
equassion
équateur
equateur
équatorial
equatorial
équatoriale
equatoriale
équatoriaux
equatoriaux
équiangle
equiangle
équidé
equidé
équidistance
equidistance
équidistant
equidistant
équidistante
equidistante
équilatéral
equilatéral
équilatérale
equilatérale
équilatéraux
equilatéraux
équilatère
equilatère
équimoléculaire
equimoléculaire
équimultiple
equimultiple
équipartition
equipartition
équipolé
equipolé
équipollé

equipollé
équipollence
equipollence
équipollente
equipollente
équipollents
equipollents
équipotente
equipotente
équipotentiel
equipotentiel
équipotentielle
equipotentielle
équipotents
equipotents
équiprobable
equiprobable
équisétale
equisétale
équisétinée
equisétinée
ersatz
ertie
eschare
eschatologie
eschatologique
escroc
estomac
eu
eucalyptus
eucharistie
eucharistique
eudémis
eue
eues
euphuisme
eurêka
eus
exact
exarchat
exeat
exemption
exequatur
exit
expugnable
extradent
extrader
extrades
f
fa
facétie
facies
faisant
faitdiversié
faitdiversière
fâme
fan
faonne
farad
farniente
fat
favus
fedayin
feddayin
feddayine
feld
feldspath
felouque
femelle
femmelette
fiat
fibrillassion
fibrille
fibrome
fichtre
fichtrement
ficit
ficus
fiteor
fjord
flamenco
flint
flirt
flirts
flutiau
foène
foëne
foetus
foil
folennaise
fongus
foot
football
footballeur
footballeuse
footing
forceps
fortiori
fournil
frangibus
frichti
frimas
fromgi
fromton
fuchsia
fucus
fuel
fundibulum
fusil
g
gable
gadget
gageure
gagman
galetas
galop
gamosépale
garden
gars
gât
gâte
gayal
gaz
geisha
generis
genèse
genevois
genevoise
genévrié
genévrière
gentiane
gentilhomme
gentilhommière
gentilshommes
gentleman
gentlemen
géosynclinal
géosynclinale
géosynclinaux
gestapiste
gestapo
guesha
gibus
gilde
gille
gimmick
gin
ginkgo
girasol
girl
glas
glaucome
gléchome
glécome
gluten
gnance
gnant
gnante
gnète
gnome
gnomes
gnou
goal
gobeleterie
goétie
goï
golden
goy
goyave
goyavié
goyim
goym
graben
gradus
graineterie
gramen
granit
gratia
gratiole
gravos
groenlandaise
groom
groseillié
gruyère
guanaco
guano
guarani
guatémaltèque
guayen
guayenne
guenille
guenon
guérillero
guinal
guinale
guinaux
guisarme
guisarmié
gus
gym
gymkhana
gymnosophie
gymnosophiste
h
habeas
habitus
hacienda
hafnium
haggis
haïku
haïtien
haleine

halicte
halieutique
haliotide
haliple
hall
hallali
haltère
hamadryade
hamadryas
hamamélis
handball
handballeur
handballeuse
hapax
hard
harpagon
haruspice
hast
hâte
hautbois
hautboïste
havresac
hawaïen
hawaïenne
hawaiien
hawaiienne
hécatombe
hectowatt
heiduque
hélépole
hello
hélodée
hendayen
hendayenne
hendyadis
hendyadyin
henry
heptasyllabe
heptasyllabique
hérédosyphilis
hérédosyphilitique
herpès
hétérosexualité
hétérosexuel
hétérosexuelle
hetman
hévéa
hexasyllabe
hexasyllabique
hiatus
hibiscus
hidalgo
hidrosadénite
hièble
hindi
hinterland
hippie
hippodrome
hippy
hollywoodien
hollywoodienne
holmium
hombre
home
homothétie
homuncule
honneur
honoris
hoplite
horaire
hui
hum
humant
humérus
humour
humus
hurluberlu
hymen
hypnoses
hypospadias
ibis
iceberg
ichthus
ichtyosaure
ictus
idiome
idiomes
idiotes
igloo
igné
ignipuncture
ignipunctures
iguane
iguanodon
iléus
imbécillité
imbroglio
immangeable
immanquable
immanquablement
immariable
immettable
impatiemment
impedimenta
impéritie
impresario
incipit
indes
indouze
infarctus
inlandsis
interruptus
interview
interviews
inuit
inuits
iris
iritis
irrespect
isatis
isochore
isoionique
isoséiste
isosiste
j
jacobus
jaguar
jamboree
jazz
jeans
jeun
joaillié
joaillière
jogging
joug
judas
jungo
k
karman
kart
kayac
kayak
keffieh
kelvin
kentia
kentrophylle
ketchup
khamsin
kieselguhr
kieselgur
kiesérite
kilbus
kilowatt
kilowattheure
kit
kiwi
knout
koumis
koumys
kouros
krach
kraft
kronprinz
kugelhof
kumquat
kvas
kwas
kyrie
l
labadens
lactosérum
lad
ladies
lady
laïus
lamdé
lamfé
laotien
lapis
laps
lapsus
lasse
latifundia
latifundium
lazzarone
lebel
ledit
leitmotiv
lepidosiren
lépidosirène
lesdites
lesdits
lexis
libris
lichen
lied
lieuxdits
lilas
lilliputien
liparis
lipome
liposoluble
lithotritie
llaise
loden
loess
loft
logos
lolo
look
looping
losange

losangé
lotus
loucherbem
loup
louveterie
lumbago
lumen
lumpen
lumpenproletariat
lunch
lupus
luth
lychnis
lys
m
machaon
machisme
macho
macroure
macroures
maelström
magenta
magnat
magnificat
magnum
maïs
maïzena
mahjong
malachite
malus
mammouth
mamselle
mamzelle
mancenillié
mangeure
maous
marc
marengo
marguillié
marqueterie
martien
mas
mat
mât
mâte
math
mathieu
maxille
maya
mayonnaise
mazout
mechta
médius
meeting
mégabit
mégalopolis
mégohm
mégohmmètre
memento
mendélévium
mendélien
mendélienne
mendélisme
menthol
mentholé
mentor
menuet
merdas
mérinos
mesdemoiselles
messieurs
métempsychose
métis
métrobus
métropolis
mezzanine
microbus
microhm
microhms
millet
milliampère
milliampèremètre
milliwatt
minibus
minimes
minutie
mirabilis
miserere
misses
moderato
moderatos
modus
moere
moëre
moeurs
monoecie
monoecies
monôme
monômes
mooniste
morbus
mordicus
mortibus
motus
mourre
mousqueterie
moyenâgeux
moyenâgeuse
mucus
muezzin
muffin
muleta
munichois
munichoise
muros
mutandis
mutatis
mycosis
myosis
myosotis
n
naos
nat
nautilus
négus
nemrod
nenni
népenthès
nerf
net
neufchâtel
nicaraguayen
nicaraguayenne
nicotiane
nietzschéen
nietzschéenne
nimbus
ninas
nobelisable
nobelium
nodus
noix
nombril
nonos
nounours
nouure
nucléus
numerus
nurse
nursery
nystagmus
oaristis
oasis
obit
obliquité
obs
occiput
octosyllabe
octosyllabique
oculus
oenothera
oersted
oerstite
oestrus
offset
oh
ohm
oigne
oille
olibrius
olisbos
omnibus
omniums
onzième
onzièmement
ophrys
opus
orchis
orémus
oribus
orthoses
otage
otalgie
otarie
otique
otite
otons
oued
oui
ouistiti
ours
out
outil
outlaw
ovibos
oxydent
oxyder
oxydes
oxyure
oxyures
oyais
oyait
oyant
oyat
oye
oyez
oyiez
oyions
oyons
p
pagaïe
paille

paléosol
pâme
pancréas
paneterie
paonne
paonneau
papaye
papayers
papeterie
papillaire
papillifère
papilliforme
papillome
papyrus
paradent
parader
parades
paradis
parafer
pardessus
parisyllabe
parisyllabique
parkinson
paros
parqueterie
pastis
pat
pâte
pathognomonique
pathos
patiemment
pechblende
pecus
pedibus
pédibus
pedigree
pelagos
pelleterie
pelvis
penalties
penalty
pénicille
pénicillé
pénicilline
pénis
pensum
penta
pentédécagone
penthiobarbital
penthode
penthotal
pentode
péplos
peppermint
péréquassion
pergélisol
périarthrite
péripétie
périsystole
périurbain
périurbaine
persil
peso
peton
peuhl
peuhle
peul
peule
peulven
peupleraie
peupleraies
peuplié
peuprès
phallus
phébus
philodendron
philosémite
philosémitisme
philosioniste
philosoviétique
philosoviétiques
phimosis
phot
phtiriasis
phtirius
physalis
picholine
pickpocket
picris
picrocholine
pidgin
piloselle
pipit
pitpit
pityriasis
pizza
pizzeria
pizzicato
placebo
placenta
placentaire
placentassion
plaid
plenum
plésiosaure
plexiglas
plexus
poix
polaroïd
polenta
pollen
polyandres
polyèdre
polyèdres
poncho
pool
posteriori
pouls
praxis
préparent
presbytie
primatie
primesaut
primesautié
primesautière
princeps
processus
prodrome
prolapsus
prompteur
pronaos
pronunciamiento
prophétie
propitiassion
propitiatoire
propolis
proscenium
prosecteur
prosectorat
prosélyte
prosélytisme
prospectus
prunus
prurit
pschents
psitt
psoas
psoriasis
psycho
psychose
psychoses
ptérosaurien
pubis
pudding
pugnace
pugnacité
puncheur
puncture
punk
pupillaire
pupillarité
pusillanime
pusillanimité
putsch
putschiste
puzzle
pyrosis
q
quadrant
quadrants
quadri
quadriceps
quadriennal
quadriennale
quadriennales
quadriennaux
quadru
quakeresse
quakerisme
quanta
quantique
quantum
quark
quartette
quartidi
quarto
quassia
quassié
quassine
quatrillion
quatuor
questeur
questure
quetsche
quia
quibus
quidam
quiddité
quiétisme
quiétiste
quillié
quincaillié
quincaillière
quinquagénaire
quinquagésime
quinquennal
quinquennale
quinquennat
quinquennaux
quintette
quintidi
quintillion

quinto
quirite
quiscale
quitus
quorum
quota
quotients
r
raccroc
rachis
racing
racket
radius
radoub
raid
rail
raïs
ramdam
raminagrobis
ranch
raout
rapidos
rapt
rasibus
raz
reaganien
reaganienne
rebours
rébus
recette
recettes
recordman
recordmen
recordwoman
recordwomen
recroqueville
recroquevilles
recul
reculs
reddition
réexamen
référendum
referment
reflex
refuge
refuges
refus
reggae
reichsmark
reichstag
reis
relatent
relation
religion
religions
relis
relut
remake
remakes
remède
remèdes
remords
remorque
remorques
renard
renards
renne
rennes
renotions
replacer
reportions
reps
requiem
réquisit
rescapé
rescapée
rescapées
rescapés
rescision
rescisions
rescisoire
rescisoires
rescousse
rescousses
rescription
rescriptions
rescrit
rescrits
résingle
résipiscence
respects
ressort
ressorts
ressource
ressources
ressuiez
ressuions
restiez
restions
restons
restoroute
restoroutes
restringente
restringentes
restringents
resu
revanche
revanches
revis
rhésus
rhétien
rhinocéros
rhizome
rhododendron
rhodoïd
rhumb
rhumé
rhument
rhumer
rhumerie
ria
riai
riant
riants
riante
rictus
rieur
rieuse
riez
riiez
riions
rimaye
rinforzando
rions
rococo
roentgen
rollmops
romsteck
röntgen
round
rouennaise
rumb
rumba
rumen
rumsteck
ruolz
rush
rushes
s
sabayon
samizdat
sanctus
sandow
sandwich
sandwiches
sangsue
sanguinité
sanguisorbe
saoul
saouls
sapiens
sapotillié
sarcome
saros
sas
satiété
satisfecit
satyriasis
scenium
schelling
scherzando
scherzo
scherzos
schibboleth
schiedam
schilling
schlich
schnaps
schnorchel
scholie
schupo
schuss
scille
scoop
scootériste
scorbut
scout
script
secundo
sécurit
seltz
sempervirens
sempervivum
sénatus
sensu
septain
septième
septièmement
septmoncel
sequin
séquoia
serpillière
sesquialtère
sesquioxyde
set
shakespearien
shakespearienne
shetland
shilling
shintô
shintoïsme
shoot

shoots
short
shunt
sialis
siemens
sigillaire
sigillé
sigillographie
sigillographique
similisent
similises
sinciput
sinus
sirop
sixain
sixième
sixièmement
skating
skunks
sleeping
sloop
slow
smart
soixantaine
soixantième
solen
sombrero
sortium
soubresaut
sourcil
soûl
soya
sparadrap
spartiate
spath
speakerine
spécimen
speculum
speech
spéos
spirillose
spleen
spot
sprint
sprints
squat
squaw
stabat
stand
staphylome
steak
stencil
stendhalien
stendhalienne
steward
stichomythie
stillassion
stillatoire
stilligoutte
stimulus
stinct
stochastique
stoechiométrie
stoechiométrique
stradivarius
stratus
strontiane
strontium
stuka
succinct
succincte
succinctement
sufficit
summum
sunlight
superman
supermen
suprématie
surf
suspect
swahili
swastika
sweepstake
swing
sycosis
syllabus
synchondrose
syndrome
syngnathe
synopsis
syphilis
tabes
tabloïd
tacet
tachine
tachistoscope
tachistoscopique
tachiscopique
tagliatelle
tahitien
talmud
tamaris
tansad
tapecul
taricheute
tas
tâte
tayaut
tease
teaseuse
teeshirt
télexiste
telligentsia
telligentzia
tentiard
tentiarde
tepidarium
tequila
térit
terminus
tertio
tétanos
tétrarchat
tétrasyllabe
tétrasyllabique
thalamus
thanatos
théosophe
théosophie
thermos
tholos
thrips
thrombus
thuya
thymus
tie
toast
toasteur
tomahawk
tonus
tophus
tournesol
tournevis
trachée
trachome
trachomes
tractus
tragus
tramway
trans
transat
transats
transept
transepts
transi
transit
transits
transportions
tremens
trépas
tribunitien
triceps
tricératops
trichiasis
trichiez
trichions
trichome
trichomes
trichomonas
trichons
trichrome
trillion
trismus
trisoc
triumviral
triumvirale
triumvirat
triumviraux
trochaïque
trochantérien
trochantérienne
trochile
trochilidé
troïk
trolleybus
truisme
trypanosome
tub
tumulus
turbith
turneps
tussuscepssion
tweed
twill
twist
typhus
tzar
tzarine
tzigane
ubiquiste
ubiquité
uhlan
unciforme
unciné
unguéal
unguéale
unguéaux
unguifère
unguis
uniaxe
uniaxé
unifier

uniovulé
upas
urêus
urolagnie
urolagnique
uruguayen
uruguayenne
urus
ut
utérus
v
valgus
vamp
vanillié
varech
varus
vasistas
vaudevillesque
vaudevilliste
vaurien
vaurienne
velarium
veld
velu
velum
veniat
vénitien
vénus
vergeure
vermis
vermout
vermouth
versus
verticille
verticillé
veto
veule
veulerie
vidimus
vindas
vitae
vitellus
vivendi
vivisecssion
vocero
voïvodat
voïvode
voïvodie
volubilis
vosgien
vosgienne
w
wagnérien
wagnérienne
walkover
wallaby
wallace
wallingant
wallon
wallonne
wallonnie
wallonnisme
wapiti
wassingue
watt
wattheure
wattman
wattmètre
wergeld
west
western
whisky
whist
wigwam
witloof
wombat
x
xérès
xérus
yacht
yachting
yachtman
yachtmen
yack
yak
yankee
yankie
yaourt
yaourtière
yard
yatagan
yearling
yen
yeoman
yeomen
yéyé
yiddish
yod
yoga
yoghourt
yogi
yogourt
yole
yougoslave
youp
youpin
youpine
youtre
youyou
yoyo
yucca
zagaie
zen
zende
zénith
zéta
ziggourat
zoo
zoom
zuchette
zurichois
zurichoise
zut
zwanze

GROUP FOUR: PROPER NAMES

1. The 816 words in this group belong only to the category of proper names and they are words which would not be transcribed correctly by our transcription rules. When a word from Group Four occurs in the input string, the phonemic transcription is copied and the feature "prnn" is set (if not already set in Step 8, page 18).

The other groups in this lexicon contain words which can be proper names as well as common nouns (e.g., adonis, boucher, boulanger, calvados, septmoncel, vénus). When used as proper names, those words —like the proper names which are transcribed correctly by our rules— are marked proper names only when they meet the conditions described in Step 8.

2. When several pronunciations exist for the same word, our computer program uses the one which appears to be the most frequent (e.g., /sɛ̃twɛ̃/ rather than /sɛ̃twɑ̃/ for Saint-Ouen). Some last names pose a special problem since there is no way to predict how the members of a family want their last name to be pronounced; here again, we have adopted the pronunciation which is the most often encountered (e.g., we transcribe

Maeterlink as /mɑtɛrlɛ̃k/ since practically all the texts in which this name is found refer to the famous writer and since this is how he pronounced his name).

3. It is impossible to list all the words which exist as proper names and as common nouns with different pronunciations and for which, in principle, a linguistic/semantic analysis should be performed. Newspaper headlines such as these:

> Suez et maigrissez sans danger!
> Singer abandonne le marché français.
> Eu égard aux problèmes industriels, Paris propose un accord.

are certainly possible, but a careful reading of Le Figaro and France-Soir during the June 1 - August 15, 1983 period yielded only one ambiguous case (the often-used phrase "LOT A VENDRE" in the classified ads section). Still, as a linguistic exercise rather than as a requirement to achieve the accuracy level set for our program, we have included Condom, Damas, Eu, Jean, Job, Lot, Rodez, Singer, Suez in our list of ambiguous spellings and we have placed lion de Saint-Marc, place Saint-Marc, Mers El-Kébir, Michel-Ange, Monte-Carlo, Monte-Cristo, Ris-Orangis on our idiom list.

Aar
Aaron
Abd
Abencérages
Aberdeen
Abner
Abraracourcix
Achéron
Achille
Adam
Addis
Aden
Adenauer
Adler
Agen
Ahmed
Aisne
Aix
Ajax
Albertville
Aléoutiennes
Alfortville
Alfred
Algésiras
Alighieri
Alix
Amadis
Amsterdam
Andropov
Anouilh
Anschluss
Antilles
Anvers
Apis
Aramis
Arras
Arvers
Asad
Asnières
Assas
Assurancetourix
Astérix
Asuncion
Auber
Aubervilliers
Auer
Augias
Aulaye
Aunis
Auschwitz
Auxerre
Auxonne
Avesnes
Azaïs
Bacchus
Bagdad
Balthus
Banyuls
Baucis
Bayreuth
Béatrix
Beaumesnil
Booz
Beauséjour
Beausoleil
Becker
Becquer
Beethoven
Begin
Belzébuth
Bengale
Benoist
Béotie
Bernhardt
Bernstein
Besnard
Bethléem
Beyrouth
Bilitis
Birmingham
Biscaye
Blas
Blum
Blumenthal
Boers
Boétie
Boieldieu
Boisgaillard
Boisguilbert
Boishardy
Boislevin
Boisrobert
Boisvin
Boris
Botzaris
Bouthillier
Brahms
Brassens
Brejnev
Bremond
Brenner
Brieuc
Brinvilliers
Brisach
Britannicus
Broglie
Bruckner
Brunehaut
Brutus
Bruxelles
Buchenwald
Burger
Burgos
Cadix
Caen
Caïn
Caracas
Carlos
Carmen
Cayenne
Celsius
Cervantes
Chaldée
Champaigne

Champdivers
Champmeslé
Chanaan
ChanghaÏ
Chantecler
Charybde
Chasles
Chastellain
Chaucer
Chaumeix
Chéops
Chesnay
Chloé
Churchill
Cid
Clemenceau
Clercq
Clipperton
Clovis
Cohen
Coislin
Confucius
Conrad
Coolidge
Copenhague
Cosme
Couderc
Couesnon
Courmayeur
Craon
Croatie
Cromwell
Crosnier
Croy
Cyrus
Dachau
Dallas
Dalmatie
Damis
Danaos
Daphnis
Darius
Darwin
Daumesnil
David
Davoust
Defferre
Delbos
Delhi
Delille
Delisle
Delmas
Déméter
Denver
Desaix
Desaleux
Desbordes
Descamps
Descartes
Deschamps
Deschanel
Descombes
Desfontaines
Desforges
Desfossés
Desgraupes
Desgrieux
Deshayes
Deshoulières
Desjardins
Deslandres
Desmarais
Desmarets
Desmichels
Desmoulins
Desnos
Desnoyers
Desormeaux
Desormes
Desplaines
Desportes
Despréaux
Desproges
Desrochers
Desrosiers
Destouches
Desvallières
Desvaux
Devos
Diafoirus
Dickens
Dieuleveult
Dioclétien
Dionysos
Disraeli
Dombasle
Domitien
Dommartin
Donatien
Doubs
Doumer
Dreyfus
Drucker
Dubost
Duchesne
Duchesnois
Duclerc
Dufresne
Dufresnoy
Dukas
Dumesnil
Dupleix
Duquesne
Duras
Dusaulx
Duseigneur
Ecuador
Eduens
Éduens
Einstein
Eisenhower
Elisabeth
Élisabeth
Elizabeth
Emmanuel
Emmanuelle
Emmaüs
Enguien
Esperey
Estaing
Esther
Estienne
Eure
Fabius
Fahrenheit
Fayard
Fayette
Febvre
Félix
Feyder
Fillioud
Fiterman
Flers
Fleurus
Florent
Foch
Fontvieille
Ford
Forest
Fos
Francis
Fred
Fréjus
Fresnay
Fresnaye
Fresne
Freud
Friedland
Galapagos
Gallieni
Galliera
Garros
Gatien
Gaudens
Gaultier
Gennevilliers
Guislain
Guislaine
Gilles
Gilliéron
Gillot
Gizeh
Gizèh
Glasgow
Goebbels
Goering
Goethe
Goliath
Gottfried
Gougenheim
Goya
Goyon
Graal
Gracchus
Grandcamp
Grandgousier
Grandjean
Grandmesnil
Grandpré
Grandval
Grandville
Grandvilliers
Gratien
Greco
Grieg
Groenland
Grosbois
Grospierre
Guatemala
Guebwiller
Guesclin
Guilhem
Gulliver
Gutenberg
Guth
Guynemer
Haarlem
Habsbourg
HaÏti
Hamlet
Harcouët
Hautmont

Havas
Hawaii
Hawaï
Haÿ
Haydn
Haye
Heidegger
Heidelberg
Heine
Heineken
Helder
Helsinki
Helvétie
Helvétius
Hemingway
Hendaye
Henriette
Hérault
Heredia
Hindenburg
Hitler
Hoffman
Hohenzollern
Hollywood
Honduras
Honegger
Hong
Hospital
Houston
Huntziger
Huygens
Huysmans
Ibsen
Illiers
Indosuez
Ingrid
Isle
Istanbul
Jackson
Jansénius
Janus
Jefferson
Jemappes
Jéricho
Joachim
Johannesburg
Johnson
Jonas
Judith
Juillac
Juliénas
Julius
Jumilhac
Jupiter
Jutland
Kader
Kansas
Kant
Kennedy
Kepler
Kerguélen
Kessler
Khrouchtchev
Kierkegaard
Kinshasa
Kléber
Knesset
Koenig
Koenigsberg
Koestler
Kominform
Komintern
Kong
Kruger
Laboulaye
Ladislas
Laghouat
Lagos
Lamennais
Lang
Laon
Laos
Largillière
Lasalle
Laurent
Laurentie
Lavater
Laye
Leclerc
Leclercq
Lefebvre
Leforest
Legentil
Lemaistre
Lenepveu
Leningrad
Lens
Lentulus
Léonidas
Léopold
Léopoldville
Leprévost
Lesage
Lesaint
Lesbos
Lesieur
Lesseps
Lesueur
Lesurques
Lévesque
Lichtenberger
Liebig
Liechtenstein
Liepvre
Likud
Lilliput
Limonest
Lincoln
Lindbergh
Linder
Lisle
Liszt
Liverpool
Locmariaquer
Lohengrin
Loing
Lombez
Longchamp
Longemer
Longjumeau
Longpérier
Longpont
Longpré
Longwy
Lorient
Loth
Loyola
Lucifer
Lucullus
Luther
Mabille
Macbeth
MacDonald
Mach
Mackenzie
Madison
Madrid
Maeterlinck
Maggi
Maixent
Malbrough
Malesherbes
Malthus
Mamers
Manchester
Manhattan
Marcellin
Marcelline
Marcus
Marius
Mathias
Matthias
Maurras
Maxwell
Mayence
Mayenne
Mayer
Mayerling
Mayet
Mayol
Mayotte
McDonald
Meaulnes
Médicis
Meilhac
Meilhan
Mékong
Melchior
Melilla
Memphis
Mendel
Mendelssohn
Mendès
Menehould
Ménélas
Mennais
Mercedes
Meslier
Mesmer
Mesnardière
Mesnières
Mesnil
Messiaen
Metternich
Metz
Meung
Meyer
Meyerbeer
Meyrueis
Michaelis
Micromégas
Midas
Milhau
Milhaud
Millau
Millevoye
Milly
Milwaukee
Minneapolis
Minos
Miramas
Miromesnil

Modigliani
Moët
Mohammed
Molotov
Mons
Montchrestien
Montevideo
Montpellier
Montpezat
Montréal
Montredon
Montrose
Montrouge
Montvilliers
Moon
Moréas
Morris
Mortimer
Muller
Müller
Munich
Murger
Nabuchodonosor
Nasser
Natchez
Nazareth
Necker
Nehru
Neper
Nesle
Nessos
Nessus
Neufchâteau
Nevers
Nicaragua
Niémen
Nietzsche
Niger
Nigeria
Nixon
Nostradamus
Nucingen
Nungesser
Obélix
Oder
Offenbach
Omer
Orléansville
Orthez
Orvilliers
Ottawa
Ouistreham
Oxford
Oyonnax
Palavas
Pallas
Panhard
Panoramix
Paraguay
Pardaillhac
Parentis
Pâris
Pasdeloup
Pauillac
Paul
Paulhan
Payot
Pelléas
Pen
Penmarch
Pensylvanie
Penthièvre
Périchole
Perros
Perthus
Petitjean
Petitpierre
Pézenas
Phidias
Picpus
Picrochole
Piedfort
Pierrelaye
Plessisville
Pleven
Ploemeur
Poe
Pollux
Polyeucte
Pons
Pontchartrain
Pontchâteau
Pontgibaud
Pontmartin
Poperen
Popocatépetl
Poros
Porthos
Portland
Poulenc
Praslin
Presles
Pretoria
Prévost
Prévostin
Prosper
Provost
Puisaye
Puylaurens
Puymorens
Pyrrhus
Quesnay
Quesnel
Quesnoy
Quimper
Quirinal
Ramsès
Raon
Reagan
Régis
Reich
Reims
Rembrandt
Rémus
Retournemer
Retz
Reuter
Rhétie
Ribeauvillé
Richter
Riom
Risle
Ritter
Robertville
Romulus
Roosevelt
Rops
Rosny
Rothschild
Rotterdam
Rouffach
Rouher
Rubens
Ruhr
Ruth
Sachs
Sacramento
Saëns
Saïd
Samoëns
Sand
Santander
Santos
Saône
Saosnois
Sarajevo
Saulnier
Schiller
Schmidt
Schneider
Schubert
Schuman
Schweitzer
Seattle
Séez
Séguers
Seignobos
Sémiramis
Senlis
Sever
Shakespeare
Shanghaï
Shannon
Shultz
Siegfried
Sieyès
Solesmes
Solferino
Spartacus
Staël
Stalingrad
Stanislas
Stendhal
Street
Suffren
Suresnes
Taiwan
Tchad
Tchernenko
Teilhard
Terpsichore
Tesnières
Tet
Têt
Texas
Thatcher
Thiers
Thorez
Thuillier
Tiflis
Timgad
Titien
Tlemcen
Tolstoï
Tourcoing
Transylvanie
Tremblaye
Trémoille
Troisgros
Troyes
Truman
Tucson

Tunis	Vélasquez	Vincent	Winnipeg
Ulrich	Venceslas	Vosges	Winston
Uranus	Vendryes	Wagner	Wonder
Uruguay	Vénétie	Wall	Xavier
Utrecht	Venezuela	Wallonie	Yahveh
Vadius	Vercingétorix	Warens	Yrieix
Vallauris	Verhaeren	Washington	Ys
Vals	Vermeer	Waterloo	Yser
Van	Vetter	Wehrmacht	Yseult
Vancouver	Viet	Werther	Zacharie
Vaulx	Villehardouin	Wilde	Zeus
Vauquer	Villers	Wilson	Ziegler
Vayer	Villiers	Wimbledon	Zürich
Vecchio	Villon	Windsor	Zurich

GROUP FIVE: SPELLINGS WHICH HAVE MORE THAN ONE PRONUNCIATION

A. General comments:

In most cases, the words are ambiguous whether they are written with upper-case or lower-case letters (e.g., sens). However, some words are ambiguous only when printed with an initial capital letter (e.g., Lot) or printed entirely with capital letters (e.g., LOT); they are not ambiguous when typed with lower-case letters (e.g., lot); such words appear in the lists below with an initial capital letter. Some words (e.g., job) have two pronunciations when they are written with lower-case letters and three pronunciations when they have capital letters; two separate algorithms are needed for these words.

B. Lists of ambiguous spellings (158 items):

1. There are 22 items (25 if we count the three words spelled with initial é and e) which are either third person plural verbs or which belong to the fuzzy classes of nouns, adjectives, and adjectives/nouns. In group a, we place the words which are normally used as adjectives (e.g., un excellent homme, with /t/ of liaison) and those which like violent are unlikely —when used as nouns— to be followed by an adjective. (C'est un violent poses no problems, but C'est un violent agressif —if it did occur— would result in a wrong liaison between the two words marked "adjective".). In group b, we place the words which normally function as nouns (e.g., couvent) plus those which —when functioning as adjectives— normally follow the noun: un produit détergent. (We do not want to classify détergent as an adjective since this would give us a wrong liaison in un détergent admirable.)

Group a:

coïncident	kɔɛ̃sidɑ̃/kɔɛ̃sidɑ́	ambi,adj,m,sg,v,pl
content	kɔ̃tɑ̃/kɔ̃tɑ́	ambi,adj,m,sg,v,pl
convergent	kɔ̃vɛrʒɑ̃/kɔ̃vɛrʒɑ́	ambi,adj,m,sg,v,pl
divergent	divɛrʒɑ̃/divɛrʒɑ́	ambi,adj,m,sg,v,pl

émergent	emɛrʒɑ̃/emɛrʒ∅	ambi,adj,m,sg,v,pl
emergent	emɛrʒɑ̃/emɛrʒ∅	ambi,adj,m,sg,v,pl
évident	evidɑ̃/evid∅	ambi,adj,m,sg,v,pl
evident	evidɑ̃/evid∅	ambi,adj,m,sg,v,pl
excellent	ɛksɛlɑ̃/ɛksɛl∅	ambi,adj,m,sg,v,pl
influent	ɛ̃flyɑ̃/ɛ̃fly	ambi,adj,m,sg,v,pl
somnolent	sɔmnɔlɑ̃/sɔmnɔl∅	ambi,adj,m,sg,v,pl
violent	vjɔlɑ̃/vjɔl∅	ambi,adj,m,sg,v,pl

Group b:

affluent	aflyɑ̃/afly	ambi,nn,m,sg,v,pl
confluent	kɔ̃flyɑ̃/kɔ̃fly	ambi,nn,m,sg,v,pl
couvent	kuvɑ̃/kuv∅	ambi,nn,m,sg,v,pl
détergent	detɛrʒɑ̃/detɛrʒ∅	ambi,nn,m,sg,v,pl
équivalent	ekivalɑ̃/ekival∅	ambi,nn,m,sg,v,pl
equivalent	ekivalɑ̃/ekival∅	ambi,nn,m,sg,v,pl
expédient	ɛkspedjɑ̃/ɛkspedi	ambi,nn,m,sg,v,pl
ferment	fɛrmɑ̃/fɛrm∅	ambi,nn,m,sg,v,pl
insolent	ɛ̃sɔlɑ̃/ɛ̃sɔl∅	ambi,nn,m,sg,v,pl
négligent	negliʒɑ̃/negliʒ∅	ambi,nn,m,sg,v,pl
parent	parɑ̃/par∅	ambi,nn,m,sg,v,pl
président	prezidɑ̃/prezid∅	ambi,nn,m,sg,v,pl
résident	rezidɑ̃/rezid∅	ambi,nn,m,sg,v,pl

2. The following verbs are either third person singular or third person plural. The same algorithm can be used to resolve the ambiguity:

convient	kɔ̃vjɛ̃/kɔ̃vi	ambi,v,sg,pl
obvient	ɔbvjɛ̃/ɔbvi	ambi,v,sg,pl
pressent	presɑ̃/prɛs∅	ambi,v,sg,pl

3. The following items are either first person plural verbs or feminine plural nouns. The same algorithm can be used to resolve the ambiguity:

acceptions	aksɛpsjɔ̃/aksɛptjɔ̃	ambi,nn,f,pl,v,pl
adoptions	adɔpsjɔ̃/adɔptjɔ̃	ambi,nn,f,pl,v,pl
affections	afɛksjɔ̃/afɛktjɔ̃	ambi,nn,f,pl,v,pl
attentions	atɑ̃sjɔ̃/atɑ̃tjɔ̃	ambi,nn,f,pl,v,pl
concoctions	kɔ̃kɔksjɔ̃/kɔ̃kɔktjɔ̃	ambi,nn,f,pl,v,pl
contentions	kɔ̃tɑ̃sjɔ̃/kɔ̃tɑ̃tjɔ̃	ambi,nn,f,pl,v,pl
contractions	kɔ̃traksjɔ̃/kɔ̃traktjɔ̃	ambi,nn,f,pl,v,pl
désertions	dezɛrsjɔ̃/dezɛrtjɔ̃	ambi,nn,f,pl,v,pl
détractions	detraksjɔ̃/detraktjɔ̃	ambi,nn,f,pl,v,pl
dictions	diksjɔ̃/diktjɔ̃	ambi,nn,f,pl,v,pl
diffractions	difraksjɔ̃/difraktjɔ̃	ambi,nn,f,pl,v,pl
éditions	edisjɔ̃/editjɔ̃	ambi,nn,f,pl,v,pl
editions	edisjɔ̃/editjɔ̃	ambi,nn,f,pl,v,pl
électrocutions	elɛktrɔkysjɔ̃/elɛktrɔkytjɔ̃	ambi,nn,f,pl,v,pl
electrocutions	elɛktrɔkysjɔ̃/elɛktrɔkytjɔ̃	ambi,nn,f,pl,v,pl
exceptions	ɛksɛpsjɔ̃/ɛksɛptjɔ̃	ambi,nn,f,pl,v,pl
excrétions	ɛkskresjɔ̃/ɛkskretjɔ̃	ambi,nn,f,pl,v,pl
exécutions	ɛgzekysjɔ̃/ɛgzekytjɔ̃	ambi,nn,f,pl,v,pl

exemptions	ɛgzɑ̃psj3/ɛgzɑ̃tj3	ambi,nn,f,pl,v,pl
infections	ɛ̃fɛksj3/ɛ̃fɛktj3	ambi,nn,f,pl,v,pl
injections	ɛ̃ʒɛksj3/ɛ̃ʒɛktj3	ambi,nn,f,pl,v,pl
inspections	ɛ̃spɛksj3/ɛ̃spɛktj3	ambi,nn,f,pl,v,pl
intentions	ɛ̃tɑ̃sj3/ɛ̃tɑ̃tj3	ambi,nn,f,pl,v,pl
interceptions	ɛ̃tɛrsɛpsj3/ɛ̃tɛrsɛptj3	ambi,nn,f,pl,v,pl
interjections	ɛ̃tɛrʒɛksj3/ɛ̃tɛrʒɛktj3	ambi,nn,f,pl,v,pl
inventions	ɛ̃vɑ̃sj3/ɛ̃vɑ̃tj3	ambi,nn,f,pl,v,pl
mentions	mɑ̃sj3/mɑ̃tj3	ambi,nn,f,pl,v,pl
notions	nosj3/nɔtj3	ambi,nn,f,pl,v,pl
objections	ɔbʒɛksj3/ɔbʒɛktj3	ambi,nn,f,pl,v,pl
options	ɔpsj3/ɔptj3	ambi,nn,f,pl,v,pl
persécutions	pɛrsekysj3/pɛrsekytj3	ambi,nn,f,pl,v,pl
portions	pɔrsj3/pɔrtj3	ambi,nn,f,pl,v,pl
rations	rasj3/ratj3	ambi,nn,f,pl,v,pl
rééditions	reedisj3/reeditj3	ambi,nn,f,pl,v,pl
réfractions	refraksj3/refraktj3	ambi,nn,f,pl,v,pl
réinventions	reɛ̃vɑ̃sj3/reɛ̃vɑ̃tj3	ambi,nn,f,pl,v,pl
relations	rølasj3/rəlatj3	ambi,nn,f,pl,v,pl
rétractions	retraksj3/retraktj3	ambi,nn,f,pl,v,pl
sécrétions	sekresj3/sekretj3	ambi,nn,f,pl,v,pl
transitions	trɑ̃zisj3/trɑ̃zitj3	ambi,nn,f,pl,v,pl
translations	trɑ̃slasj3/trɑ̃slatj3	ambi,nn,f,pl,v,pl

4. The following items are masculine singular nouns or infinitives. The same algorithm can be used to resolve the ambiguity. The first group has the contrast /e/ - /ɛr/; the second group has the contrast /e/ - /œr/:

bitter	bitɛr/bite	ambi,nn,m,sg,inf
boxer	bɔksɛr/bɔkse	ambi,nn,m,sg,inf
carter	karter/karte	ambi,nn,m,sg,inf
corner	kɔrnɛr/kɔrne	ambi,nn,m,sg,inf
ester	ɛstɛr/ɛste	ambi,nn,m,sg,inf
manager	manadʒɛr/manadʒe	ambi,nn,m,sg,inf
mater	matɛr/mate	ambi,nn,f,sg,inf
palmer	palmɛr/palme	ambi,nn,m,sg,inf
placer	plasɛr/plase	ambi,nn,m,sg,inf
porter	pɔrtɛr/pɔrte	ambi,nn,m,sg,inf
poster	pɔstɛr/pɔste	ambi,nn,m,sg,inf
reporter	røpɔrtɛr/rəpɔrte	ambi,nn,m,sg,inf
super	sypɛr/sype	ambi,nn,m,sg,inf
supporter	sypɔrtɛr/sypɔrte	ambi,nn,m,sg,inf
flipper	flipœr/flipe	ambi,nn,m,sg,inf
interviewer	ɛ̃tɛrvjuvœr/ɛ̃tɛrvjuve	ambi,nn,m,sg,inf
sprinter	sprintœr/sprinte	ambi,nn,m,sg,inf
squatter	skwatœr/skwate	ambi,nn,m,sg,inf
stripper	stripœr/stripe	ambi,nn,m,sg,inf

5. The following items are either masculine singular adjectives or infinitives. The same algorithm can be used to resolve the ambiguity:

cacher	kaʃɛr/kaʃe	ambi,adj,ag,an,inf

The word fier is on the idiom list.

6. The following words are either nouns or verbs (first and second persons singular). The same algorithm can be used to resolve the ambiguity:

bois	bwɑ/bwa	ambi,nn,m,an,v,sg
bus	bys/by	ambi,nn,m,an,vbpastpr
lis	lis/li	ambi,nn,m,an,v,sg

The words sens and vis are on the idiom list.

7. The following words are third person singular verbs or masculine singular nouns. The same algorithm can be used to resolve the ambiguity.

lut	lyt/ly	ambi,nn,m,sg,v,sg
rit	rit/ri	ambi,nn,m,sg,v,sg

8. The following words are either verbs (second person singular) or nouns. The same algorithm can be used to resolve the ambiguity:

lias	ljɑs/lja	ambi,nn,m,an,v,sg
pagayes	pagaj/pagɛj	ambi,nn,f,pl,v,sg
rhumes	rymɸ/rɔmɸ	ambi,nn,m,pl,v,sg
trustes	trystɸ/trœstɸ	ambi,nn,f,pl,v,sg

9. The following words are either verbs (first and third persons singular) or nouns. The same algorithm can be used to resolve the ambiguity:

pagaye	pagaj/pagɛj	ambi,nn,f,sg,v,sg
rhume	rymɸ/rɔmɸ	ambi,nn,m,sg,v,sg
truste	trystɸ/trœstɸ	ambi,nn,f,sg,v,sg

10. Each of the following words needs its own algorithm to resolve the ambiguity:

as	ɑs/a	ambi,nn,m,an,v,sg
auto	oto/awto	ambi,nn,f,sg,nn,m,sg
bẽta	bɛtɑ/bɛta	ambi,adjnn,m,sg,nn,m,sg
bis	bi/bis	ambi,adj,m,an,adv
But	byt/by	ambi,prnn,nn,m,sg
campos	kɑ̃po/kɑ̃pos	ambi,nn,m,an,nn,m,an
cassis	kasis/kasi	ambi,nn,m,an,nn,m,an
chut	ʃyt/ʃy	ambi,adv,v,sg
Condom	kɔ̃dɔ̃/kɔ̃dɔm	ambi,prnn,nn,m,sg
cossus	kɔsy/kɔsys	ambi,adj,m,pl,nn,m,an
Damas	dama/damɑs	ambi,nnvb,prnn
donc	dɔ̃/dɔ̃k	ambi,coord,coord
E	ɛst/e	ambi,nn,m,sg,nn,m,sg
Eu	y/ø	ambi,ptp,m,sg,prnn
exprès	ɛkspre/ɛkspreɛs	ambi,adv,adjnn,m,an
fils	fis/fil	ambi,nn,m,an,nn,m,pl
flous	flu/flus	ambi,adjnn,m,pl,nn,m,an
forte	fɔrtɸ/fɔrte	ambi,adjnn,f,sg,nnadv

gens	ʒɛ̃s/ʒɑ̃	ambi,nn,f,an,nn,sg,an
haste	ˊastɬ/astɬ	ambi,nn,f,sg,nn,f,sg
hélas	elɑs/ˊelɑ	ambi,adv,v,sg
jet	ʒɛ/dʒɛt	ambi,nn,m,sg,nn,m,sg
job	ʒɔb/dʒɔb	ambi,nn,m,sg,nn,m,sg
Job	ʒɔb/dʒɔb/ʒob	ambi,nn,m,sg,nn,m,sg,prnn
Jobs	ʒɔb/dʒɔb	ambi,nn,m,pl,nn,m,pl
lacs	lak/lɑ	ambi,nn,m,pl,nn,m,an
las	lɑ/lɑs	ambi,adj,m,an,adv
Lot	lo/lɔt	ambi,nn,m,sg,prnn
Marc	mar/mark	ambi,nn,m,sg,prnn
minerai	minre/minrɛ	ambi,v,sg,nn,m,sg
O	wɛst/o	ambi,nn,m,sg,nn,m,sg
ouïe	wi/uj	ambi,nn,f,sg,adv
papas	papa/papɑs	ambi,nn,m,pl,nn,m,an
portos	pɔrto/pɔrtos	ambi,nn,m,pl,nn,m,pl
pub	pyb/pœb	ambi,nn,f,sg,nn,m,sg
punch	pɔ̃ʃ/pœnʃ	ambi,nn,m,sg,nn,m,sg
ras	rɑ/rɑs	ambi,adjnn,m,an,nn,m,an
Rodez	rɔde/rɔdɛz	ambi,v,pl,prnn
S	syd/ɛs	ambi,nn,m,sg,nn,m,sg
seps	sɛp/sɛps	ambi,nn,m,pl,nn,m,an
Singer	sɛ̃ʒe/sɛ̃ʒɛr	ambi,inf,prnn
soit	swa/swat	ambi,v,sg,adv
soustraite	sutrɛtɬ/sustrɛtɬ	ambi,v,sg,ptp,f,sg
soustraites	sutrɛtɬ/sustrɛtɬ	ambi,v,sg,ptp,f,pl
Suez	sɥe/sɥɛz	ambi,v,pl,prnn
suspense	syspɑ̃sɬ/syspɛns	ambi,nn,f,sg,nn,m,sg
trias	trijɑs/trija	ambi,nn,m,an,v,sg

The words est, Jean, N, os, plus, tous, y are on the idiom list.

C. Words which cease to be ambiguous when a final s is added:

Only a few words remain ambiguous when a final letter s is added (e.g., auto/autos, jet/jets, suspense/suspenses). Those which are not ambiguous when s is added must be caught by our program since, otherwise, they would be marked ambiguous after the final s is removed. Words ending with ents (e.g., violents) are caught by the table of endings; minerais and the words ending with ers (except fiers which is on the list of idioms) have already been included in Group One; fortes is one of the adjectives in Group Two. The words which remain to be added to the lexicon are:

Buts	by	nn,m,pl
chuts	ʃyt	nn,m,pl
Condoms	kɔ̃dɔm	nn,m,pl
Eus	y	ptp,m,pl
Jeans	dʒin	nn,m,pl
Lots	lo	nn,m,pl
luts	lyt	nn,m,pl
Marcs	mar	nn,m,pl
ouïes	wi	nn,f,pl
rits	rit	nn,m,pl

D. Words which cease to be ambiguous when a prefix is added:

The words which cease to be ambiguous when a prefix is added have been discussed on pages 28 and 29.

GROUP SIX

Abbreviations which have the "ww" feature marker and which are not pronounced letter by letter. The linguistic feature "noun" is attached to all these 186 abbreviations.

wwabs	obɜswɛdɸ	wwcredif	kredif
wwadac	adak	wwcrɛ̃dif	kredif
wwadav	adav	wwcredoc	kredɔk
wwafat	afat	wwcreps	kreps
wwafnor	afnɔr	wwcrica	krika
wwaid	ɛd	wwcrif	krif
wwalat	alat	wwcrous	krus
wwanah	ana	wwdafco	dafko
wwanpe	anpe	wwdatar	datar
wwantiope	ɑ̃tjɔp	wwdest	dɛst
wwarrco	arko	wwdeug	døg
wwasa	aza	wwdin	din
wwassedic	asedik	wwdipas	dipas
wwbenelux	benelyks	wwdom	dɔm
wwcapes	kapɛs	wwduel	dyɛl
wwcapet	kapɛt	wwdues	dyɛs
wwceca	seka	wwecu	eky
wwcecos	sekos	wwedhec	edɛk
wwcedex	sedɛks	wwelf	ɛlf
wwceres	serɛs	wwena	ena
wwcɛ̃res	serɛs	wwerap	erap
wwcɛ̃rɛ̃s	serɛs	wwetdam	ɛtdam
wwcern	sɛrn	wwfane	fan
wwcid	sid	wwfar	far
wwcidunati	sidynati	wwfen	fɛn
wwcii	sedøzi	wwfiat	fjat
wwcleru	klery	wwfides	fidɛs
wwcnes	knɛs	wwfinul	finyl
wwcnet	knɛt	wwfnac	fnak
wwcnit	knit	wwfnef	fnɛf
wwcnous	knus	wwforma	fɔrma
wwcodefi	kɔdefi	wwgar	gar
wwcoder	kɔdɛr	wwgatt	gat
wwcodevi	kɔdevi	wwgestapo	gɛstapo
wwcodofil	kɔdɔfil	wwgi	dʒiaj
wwcon	kɔ̃	wwgifas	ʒifas
wwcofar	kɔfar	wwglam	glam
wwcogefi	kɔʒefi	wwgrece	grɛs
wwcomecom	kɔmekɔm	wwgueule	gœl
wwcominform	kɔminfɔrm	wwidate	idat
wwcopar	kɔpar	wwidhec	idɛk

wwidi	idi
wwifop	ifɔp
wwifres	ifrɛs
wwigame	igam
wwigas	igas
wwigen	igɛn
wwina	ina
wwined	inɛd
wwinfrep	infrɛp
wwinra	iŋra
wwinri	inri
wwinsee	insɛ
wwinserm	ɛ̃sɛrm
wwipes	ipɛs
wwipsos	ipsɔs
wwira	ira
wwircam	irkam
wwirep	irɛp
wwiret	irɛt
wwjac	ʒak
wwjec	ʒɛk
wwjic	ʒik
wwjoc	ʒɔk
wwlaser	lazɛr
wwlicra	likra
wwlof	lɔf
wwmaser	mazɛr
wwmerde	mɛrd
wwmidem	midɛm
wwmijid	miʒid
wwmin	min
wwnasa	naza
wwonera	ɔnera
wwonisep	ɔnisɛp
wwonu	ɔny
wwopep	ɔpɛp
wworsec	ɔrsɛk
wwotan	otɑ̃
wwotase	otaz
wwovni	ɔvni
wwpac	pak
wwpaf	paf
wwpal	pal
wwpam	pam
wwpap	pap
wwpaz	paz
wwpfat	pefat
wwpic	pik
wwputain	pytɛ̃
wwradar	radar
wwrecours	rəkur
wwsabena	sabena
wwsac	sak
wwsacem	sakɛm
wwsagi	saʒi
wwsalope	salɔp
wwsalt	solt
wwsamu	samy
wwsarl	sɔsjetearɛsp3 sabilitelimite
wwsecam	sekam
wwseita	seta
wwshape	ʃep
wwsicav	sikav
wwsicob	sikɔb
wwsid	sid
wwsida	sida
wwsimca	simka
wwsirpa	sirpa
wwsitef	sitɛf
wwsmag	smag
wwsmic	smik
wwsmig	smig
wwsmip	smip
wwsnca	snɛka
wwsnecma	snɛkma
wwsnes	snɛs
wwsnesup	snɛsyp
wwsni	sni
wwsnias	snjas
wwsofirad	sɔfirad
wwsofres	sɔfrɛs
wwsofretu	sɔfrety
wwtir	tir
wwtom	tɔm
wwtuc	tyk
wwunaf	ynaf
wwunapel	ynapɛl
wwunapi	ynapi
wwunati	ynati
wwunedic	ynedik
wwunef	ynɛf
wwunesco	ynɛsko
wwunicef	ynisɛf
wwunostra	ynɔstra
wwunra	ynra
wwurssaf	yrsaf
wwvip	vip
wwzac	zak
wwzad	zad
wwzeat	zat
wwzede	zɛd
wwziv	ziv
wwzup	zyp
wwzut	zyt

GROUP SEVEN

Abbreviations which have the "ww" marker and in which a letter E is pronounced /e/. The feature "noun" is attached to these eight abbreviations.

wwaef	aeef	wwedf	edeef	wwene	eena	wweor	eoer
wweao	eao	wwegf	egeef	wwens	eenes	wweta	etea

GROUP EIGHT

The 29 abbreviations in this group do not have intervening periods and contain small letters; they are listed here either because they would not be transcribed correctly by our rules or because they are pronounced letter by letter. The linguistic feature "noun" is attached to these abbreviations.

afat	afat	cominform	kɔminfɔrm	sfio	esefio
anpe	anpe	credif	kredif	simca	simka
assedic	asedik	credoc	kredɔk	snes	snes
benelux	benelyks	crous	krus	snesup	snesyp
cecos	sekos	ifres	ifres	sofirad	sɔfirad
cécos	sekos	insee	inse	sofres	sɔfres
cedex	sedeks	ipsos	ipsos	unedic	ynedik
ceres	seres	ratp	eratepe		
céres	seres	sabena	sabena		
cnet	knet	sacem	sakem		
cnous	knus	secam	sekam		

I D I O M S

This group (1128 items) contains two types of word combinations:

1. The first type includes some word combinations which must be treated separately because they contain one or several words which, in those combinations, do not have their usual pronunciation (e.g., chef-d'oeuvre) or because a type of liaison not ordinarily made occurs (e.g., accent aigu), or because including an expression on the list facilitates the linguistic analysis (e.g., il en est, je puis, pull-over).[17]

We only have to include the combinations which present problems. For example, guets-apens is on the list, but not guet-apens since the singular form is transcribed correctly by our regular program; likewise, vingt-neuf ans is on the list, but not trente-neuf ans since trente does not present a problem and neuf ans is already listed.

We also include some combinations in which words, although ambiguous by themselves, are no longer ambiguous (e.g., compte-fils, vice-président).

This list of idioms has been collected over the last ten years and we believe it includes practically all the combinations which are likely to occur in the types of text for which our program has been written. Any idiomatic expression (whether hyphenated or not) not found on this list is

assumed to be transcribed correctly by our transcription rules.

2. The second type includes word combinations which are function words. For example, we need to isolate combinations such as si bien que, pour que so that they can be marked as conjunctions of subordination.

Our program uses the idiom list as follows: if, during the lexicon search, a word from the string being transcribed matches a head word in the idiom list, the program looks at the following words to determine whether one of the combinations listed under the head word is present. For example, if the word pieds is found, the program looks at the following words to determine whether the expression pieds et poings liés is used. If so, the whole expression is transcribed as one string /pjezepwɛ̃lje/ and marked "adv"; if not, pieds is transcribed and marked "nn,m,pl". Words which are listed with a capital letter (e.g., Ben) are found only if the feature "cap" is set. (Therefore, a word may appear in the lexicon twice: once with a capital letter and once with a lower-case letter.) Words which are listed with a hyphen (e.g., vice-président) are found only if the features "prehyph" and "posthyph" are set.

à	prep:82,89,97,102
* bras ouverts	adv
* cause de	prep;97
* cause des	prep,z;97
* cause diy	prep:97
* cause du	prep;97
* ce dont	sub,t
* ce kiy	sub
* ce que	sub
* ce qui	sub
* condition de	prep
* condition diy	prep
* condition kiy	sub
* condition que	sub
* côté de	prep;97,112
* côté des	prep,z;97
* côté diy	prep;97
* côté du	prep;97
* défaut de	prep;97
* défaut des	prep,z;97
* défaut diy	prep;97
* défaut du	prep;97
* huis clos	adv
* liy abri de	prep;97
* liy abri des	prep,z;97
* liy abri diy	prep;97
* liy abri du	prep;97
* la faveur de	prep;97
* la faveur des	prep,z;97
* la faveur diy	prep;97
* la faveur du	prep;97
* la merci de	prep;97
* la merci des	prep,z;97
* la merci diy	prep;97
* la merci du	prep;97
* liy égard de	prep;97,112
* liy égard des	prep,z;97
* liy égard diy	prep;97
* liy égard du	prep;97
* liy heure	adv
* mesure kiy	sub
* mesure que	sub
* moins de	prep
* moins diy	prep
* moins kiy	sub
* moins que	sub
* travers	prep;112
accent	nn,m,sg
* aigu	nn,m,sg
accents	nn,m,pl
* aigus	nn,m,pl
acte	nn,m,sg
* un	nn,m,sg
ad	prep
* patres	adv
afin	prep
* de	prep;89,93
* diy	prep;89,93
* kiy	sub
* que	sub
agit	v,sg
* prop	nn,f,sg
ainsi	adv
* kiy	sub
* que	sub
alors	adv

```
* kiy                    sub
* que                    sub
après                   prep,z;89,97,102
* demain                 adv
* kiy                    sub
* que                    sub
arts                    nn,m,pl
* et métiers             nn,m,pl
assez                   adv
* de                     det
* diy                    det
attaché                 adjnnpp,m,sg
* case                   nn,m,sg
au                      det,m,sg;103
* cas où                 sub
* dedans de              prep;97,112
* dedans des             prep,z;97
* dedans diy             prep;97
* dedans du              prep;97
* dehors de              prep;97,112
* dehors des             prep,z;97
* dehors diy             prep;97
* dehors du              prep;97
* delà de                prep;97,112
* delà des               prep,z;97
* delà diy               prep;97
* delà du                prep;97
* dessous de             prep;97,112
* dessous des            prep,z;97
* dessous diy            prep;97
* dessous du             prep;97
* dessus de              prep;97,112
* dessus des             prep,z;97
* dessus diy             prep;97
* dessus du              prep;97
* devant de              prep;97,112
* devant des             prep,z;97
* devant diy             prep;97
* devant du              prep;97
* fait                   adv
* fur et à mesure kiy    sub
* fur et à mesure que    sub
* leur                   nn,m,sg
* lieu de                prep;97
* lieu des               prep,z;97
* lieu diy               prep;97
* lieu du                prep;97
* lieu kiy               sub
* lieu que               sub
* mien                   nn,m,sg
* milieu de              prep;97
* milieu des             prep,z;97
* milieu diy             prep;97
* milieu du              prep;97
* moins                  adv
* nôtre                  nn,m,sg
* revoir                 adv
* sien                   nn,m,sg
* tien                   nn,m,sg
* travers de             prep;97,112
* travers des            prep,z;97
* travers diy            prep;97
* travers du             prep;97
* vôtre                  nn,m,sg
auprès                  adv
* de                     prep;97,112
* des                    prep,z;97
* diy                    prep;97
* du                     prep;97
aussitôt                adv
* kiy                    sub
* que                    sub
autant                  adv
* de                     det
* diy                    det
* kiy                    sub
* que                    sub
autour                  adv
* de                     prep;97,112
* des                    prep,z;97
* diy                    prep;97
* du                     prep;97
aux                     det,ag,pl,z
* autres                 nn,ag,pl
* dépens de              prep;97,112
* dépens des             prep,z;97
* dépens diy             prep;97
* dépens du              prep;97
* leurs                  nn,ag,pl
* miennes                nn,f,pl
* miens                  nn,m,pl
* nôtres                 nn,ag,pl
* siennes                nn,f,pl
* siens                  nn,m,pl
* tiennes                nn,f,pl
* tiens                  nn,m,pl
* unes                   nn,f,pl
* uns                    nn,m,pl
* vôtres                 nn,ag,pl
avant                   nnprep;97,102,112
* coureur                adj,m,sg
* de                     prep;89,93
* diy                    prep;89,93
* hier                   adv
* kiy                    sub
* que                    sub
```

```
bal                 nn,m,sg
* des quatre arts   nn,m,sg
basket              nn,m,sg
* ball              nn,m,sg
basses              adj,f,pl
* alpes             nn,f,pl
beau                adjnn,m,sg
* fils              nn,m,sg
beaucoup            adv
* de                det
* des               det,z
* diy               det
Ben                 adv
* Gourion           prnn
* Hur               prnn
bien                nnadv,n
* aimé              adjnn,m,sg
* aimée             adjnn,f,sg
* aimées            adjnn,f,pl
* aimés             adjnn,m,pl
* des               det,z
* être              nn,m,sg
* kiy               sub
* que               sub
blanc               adjnn,m,sg
* estoc             nn,m,sg
* étoc              nn,m,sg
blue                adj
* jean              nn,m,sg
* jeans             nn,m,pl
boeuf               nn,m,sg
* gras              nn,m,sg
Bourg               prnn
* en Bresse         prnn
bout                nnvb
* à bout            adv
bras                nn,m,an
* et jambes         nn,m,pl
Buenos              prnn
* Aires             prnn
Ca                  prnn
* dix               prnn
call                nn
* girl              nn,f,sg
cap                 nn,m,sg
* hornier           nn,m,sg
* horniers          nn,m,pl
cas                 nn,m,an
* échéant           nn,m,sg
ce                  dp;77,94,95,96,103
* à quoi            sub
* dont              sub,t
* kiy               sub
* que               sub
* qui               sub
celui               nn,m,sg;106
* ci est            v,sg,t
* ci                nn,m,sg;106
* là est            v,sg,t
* là                nn,m,sg;106
celle               nn,f,sg;106
* ci est            v,sg,t
* ci                nn,f,sg;106
* là est            v,sg,t
* là                nn,f,sg;106
celles              nn,f,pl;104
* ci                nn,f,pl;104
* là                nn,f,pl;104
cent                nb,t
* un                nb,n
* une               nb
* unième            nb
ceux                nn,m,pl;104
* ci                nn,m,pl;104
* là                nn,m,pl;104
Ch.                 prnn
* de Gaulle         prnn
chapitre            nn,m,sg
* un                nn,m,sg
châteaux            nn,m,pl
* en Espagne        nn,m,pl
chef                nn,m,sg
* diy oeuvre        nn,m,sg
chefs               nn,m,pl
* diy oeuvre        nn,m,pl
chewing             adj
* gum               nn,m,sg
chiffre             nn,m,sg
* un                nn,m,sg
cinq                nb,tn
* Mars              prnn
cinquante           nb
* huit              nb,tn
ciy                 spr;96
* est à dire        adv
* est un fait       adv
* est               v,sg,t
comme               prep;97,102
* si                sub
* sil               sub
commedia            nn,f,sg
* delle arte        nn,f,sg
comment             adv
* allez vous ]      adv
* * * aujourdui     adv
* * * ce            det,m,sg
```

```
* * * ces          det,ag,pl
* * * cet          det,m,sg
* * * cette        det,f,sg
compte             nnvb
* fils             nn,m,an
corps              nn,m,an
* et âme           adv
* et biens         adv
croc               nn,m,sg
* en jambe         nn,m,sg
crocs              nn,m,pl
* en jambe         nn,m,pl
de                 prep;80,89,97,102
* bas en haut      adv
* bric et * broc   adv
* but en blanc     adv
* ce dont          sub,t
* ce kiy           sub
* ce que           sub
* ce qui           sub
* chez             prep;97,112
* crainte de       prep
* crainte diy      prep
* crainte kiy      sub
* crainte que      sub
* cujus            adv
* façon à          prep
* façon kiy        sub
* façon que        sub
* facto            adv
* fait             adv
* fond en comble   adv
* haut en bas      adv
* jure             adv
* liy est          adv
* mieux en mieux   adv
* moins en moins   adv
* peur de          prep
* peur diy         prep
* peur kiy         sub
* peur que         sub
* pied en cap      adv
* pis en pis       adv
* plus en plus     adv
* point en point   adv
* profundis        adv
* retard           adv
* sorte kiy        sub
* sorte que        sub
* temps à autre    adv
* temps en temps   adv
* visu             adv
dead               adj
* heat             nn,m,sg
depuis             prep;97,112
* kiy              sub
* que              sub
des                det,ag,pl,z
* autres           nn,f,pl
* leurs            nn,ag,pl
* miennes          nn,f,pl
* miens            nn,m,pl
* nôtres           nn,ag,pl
* pieds à la tête  adv
* siennes          nn,f,pl
* siens            nn,m,pl
* tiennes          nn,f,pl
* tiens            nn,m,pl
* unes             nn,f,pl
* uns              nn,m,pl
* vôtres           nn,ag,pl
dès                prep,z;97,102,112
* kiy              sub
* que              sub
deus               nn,m,sg
* ex machina       nn,m,sg
diversiy           nn,m,pl
* club             nn,m,sg
dix                nb,tn
* huit             nb,tn
* sept             nb
* neuf ans         nn,m,pl
* neuf heures      nn,f,pl
* neuf             nb
diy                prep;80,89,97,102
* autant plus kiy  sub
* autant plus que  sub
* avance           adv
* entre            prep;97,112
* ores et déjà     adv
* un bout à liy autre adv
dommages           nn,m,pl
* et intérêts      nn,m,pl
dos                nn,m,an
* à dos            adv
droit              nnadjadv
* au but           adv
du                 det,m,sg;103
* leur             nn,m,sg
* mien             nn,m,sg
* nôtre            nn,m,sg
* sien             nn,m,sg
* tien             nn,m,sg
* tout au tout     adv
* vôtre            nn,m,sg
dum                adj,ag,an
```

* dum	adjnn,ag,an
eau	nn,f,sg
* forte	nn,f,sg
eaux	nn,f,pl
* et forêts	nn,f,pl
elle	allpron,f,sg; 77,96,98,106
* en est	v,sg,t
* est	v,sg,t
en	ppr,n;79,86,88, 92,97,105,112
* avance	adv
* bas de	prep;97,112
* bas des	prep,z;97
* bas diy	prep;97
* bas du	prep;97
* chair et en os	adv
* dépit de	prep;97
* dépit des	prep,z;97
* dépit diy	prep;97
* dépit du	prep;97
* face de	prep;97,112
* face des	prep,z;97
* face diy	prep;97
* face du	prep;97
* fait	adv
* faveur de	prep;97
* faveur des	prep,z;97
* faveur diy	prep;97
* faveur du	prep;97
* haut de	prep;97,112
* haut des	prep,z;97
* haut diy	prep;97
* haut du	prep;97
* retard	adv
* temps et lieu	adv
* temps utile	adv
est	ambi,nn,m,sg,v,sg,t
* ce	v,sg
* elle	v,sg
* il	v,sg
* on	v,sg
et	coord;107
* cetera	adv
* cétera	adv
Etats	nn,m,pl
* Unis	prnn
ex	adj
* aequo	adv
* équo	adv
face	nn,f,sg
* à	prep;97,112
* au	prep;97
* aux	prep,z;97
fait	adjnnvb
* accompli	nn,m,sg
* acquis	nn,m,sg
* et cause pour	prep
faits	adjnnpp,m,pl
* et gestes	nn,m,pl
faut	v,sg
* bien kiy	v,sg
* bien que	v,sg
faute	nnvb
* de	prep;89,97
* diy	prep;89,97
femmes	nn,f,pl
* et enfants	nn,m,pl
fier	ambi,adj,m,sg,inf
* à bras	nn,m,sg
fiers	adj,m,pl
* à bras	nn,m,pl
fox	nn,m,an
* trot	nn,m,sg
franc	nnadjadv
* alleu	nn,m,sg
* arbitre	nn,m,sg
* archer	nn,m,sg
* étrier	nn,m,sg
francs	adjnn,m,pl
* alleus	nn,m,pl
* archers	nn,m,pl
garçons	nn,m,pl
* et filles	nn,m,pl
gas	nn
* oil	nn,m,sg
Germain	prnn
* liy Auxerrois	prnn
globe	nn,m,sg
* trotter	nn,m,sg
* trotters	nn,m,pl
grâce	nn,f,sg
* à	prep;97,112
* au	prep;97
* aux	prep,z;97
guets	nn,m,pl
* apens	nn,m,pl
Gulf	nn,m,sg
* Stream	nn,m,sg
hand	nn
* ball	nn,m,sg
hold	nn,m,sg
* up	nn,m,an
hommes	nn,m,pl
* et femmes	nn,m,pl
hors	prep;97,112

* de	prep;97,112
* des	prep,z;97
* diy	prep;97
* du	prep;97
* kiy	sub
* que	sub
il	spr,m,sg;106
* en est	v,sg,t
* est	v,sg,t
* était une fois	adv
* leur est	v,sg,t
* ne leur est	v,sg,t
in	prep
* absentia	adv
* abstracto	adv
* articulo mortis	adv
* dix huit	adjnn,ag,an
* douze	adjnn,ag,an
* extenso	adv
* extremis	adv
* octavo	adjnn,ag,an
* pace	adv
* partibus	adv
* petto	adv
* plano	adv
* quarto	adjnn,ag,an
* seize	adjnn,ag,an
* situ	adv
* trente deux	adjnn,ag,an
* vingt quatre	adjnn,ag,an
* vitro	adv
* vivo	adv
je	spr,sg
* puis	v,sg
* ne puis	v,sg
Jean	ambi,prnn,nn,m,sg
* foutre	nn,m,sg
* le blanc	nn,m,sg
Jésus	prnn
* Christ	prnn
jeunes	adjnn,ag,pl
* et vieux	nn,m,pl
jeux	nn,m,pl
* olympiques	nn,m,pl
jiu	nn,m,sg
* jitsu	nn,m,sg
jusquiy	prep;97
* à ce kiy	sub
* à ce que	sub
* à	prep;97
* au	prep;97
* aux	prep,z;97
* en	prep,n;97
jusques	prep,z
* et y compris	prep
knock	nn,m,sg
* out	nnadv
Kuo	nn,m,sg
* Min Tang	nn,m,sg
la	adndp,f,sg;78,88, 92,94,95,105,108
* leur	nn,f,sg;106
* mienne	nn,f,sg;106
* nôtre	nn,f,sg;106
* sienne	nn,f,sg;106
* tienne	nnvb
* vôtre	nn,f,sg;106
là	adv
* bas	adv
* dedans	adv
* dessous	adv
* dessus	adv
* haut	adv
le	dp,m,sg;78,81,88, 92,94,95,105,108
* fait est kiy	nn,m,sg
* fait est que	nn,m,sg
* huis clos	nn,m,sg
* leur	nn,m,sg;106
* mien	nn,m,sg;106
* nôtre	nn,m,sg;106
* sien	nn,m,sg;106
* tien	nn,m,sg;106
* vôtre	nn,m,sg;106
les	dp,ag,pl,z;78,81, 88,92,94,95,101
* autres	nn,ag,pl
* est	v,sg,t
* leurs	nn,ag,pl;104
* miennes	nn,f,pl;104
* miens	nn,m,pl;104
* nôtres	nn,ag,pl;104
* siennes	nn,f,pl;104
* siens	nn,m,pl;104
* tiennes	nnvb
* tiens	nnvb
* unes	nn,f,pl
* uns	nn,m,pl
* vôtres	nn,ag,pl;104
lès	prep
* Elbeuf	prnn
livre	nn,ag,sg
* un	nn,m,sg
lion	nn,m,sg
* de Saint Marc	nn,m,sg
liy	dp,ag,sg;78,88,92,

	94,95,99,105,108
* autre	nn,ag,sg
* un	nn,m,sg
* une	nn,f,sg
loin	adv
* de	prep;89,93,97
* des	prep,z;97
* diy	prep;89,93,97
* du	prep;97
* kiy	sub
* que	sub
lors	prep
* de	prep;97,112
* des	prep,z;97
* diy	prep;97
* du	prep;97
Los	det
* Angeles	prnn
lui	allpron,ag,sg;79,87, 88,91,98,100,101
* est	v,sg,t
ma	det,f,sg
* jong	nn,m,sg
made	ptp
* in	prep
mah	nn,m,sg
* jong	nn,m,sg
main	nn,f,sg
* forte	nn,f,sg
maintenant	adv
* kiy	sub
* que	sub
malgré	prep;97,102,112
* kiy	sub
* que	sub
Mao	prnn
* Tse Toung	prnn
mea	det
* culpa	nn,m,sg
mers	nn,f,pl
* El-Kébir	prnn
mesdames	nn,f,pl
* et messieurs	nn,m,pl
mezza	adj,f,sg
* voce	adv
mezzo	adj,m,sg
* tinto	nn,m,an
Michel	prnn
* Ange	prnn
mille	adjnn
* huit	nb,tn
* onze	nb
* un	nb,n
minus	adjnn,ag,an
* habens	nn,ag,an
miy	di,sg;79,84,88, 90,99,100
* as-tu-vu	nn,m,an
* as-tu-vue	nn,f,an
* est	v,sg,t
moins	adv,z
* de	det
* diy	det
monte	v,sg
* Carlo	prnn
* Cristo	prnn
monts	nn,m,pl
* et merveilles	nn,m,pl
* et par vaux	nn,m,pl
mot	nn,m,sg
* à mot	adv
* outil	nn,m,sg
muet	adjnn,m,sg
* aveugle	adjnn,m,sg
music	nn,f,sg
* hall	nn,m,sg
N	ambi,nn,m,sg, nn,m,sg
* de D	adv
Nations	nn,f,pl
* unies	prnn
ne	neg
* l'est	v,sg,t
neuf	adj
* ans	nn,m,pl
* heures	nn,f,pl
new	adj
* york	prnn
night	nn,f,sg
* club	nn,m,sg
niy	neg
* est	v,sg,t
* importe comment	adv
* importe où	adv
* importe quand	adv
* importe qui	pr,sg
* importe quoi	pr,sg
no	adj
* maniy s land	nn,m,sg
non	adv
* stop	adjnn,ag,an
nord	nn,m,sg
* africain	nn,m,sg
* africaine	nn,f,sg
* africaines	nn,f,pl
* africains	nn,m,pl

```
* américain          nn,m,sg
* américaine         nn,f,sg
* américaines        nn,f,pl
* américains         nn,m,pl
* asiatique          adj,ag,sg
* asiatiques         adj,ag,pl
* est                nn,m,sg
* ouest              nn,m,sg
nous                 allpron,ag,pl,z;79,
                     84,88,92,96,98,100
* est                v,sg,t
nuit                 nnvb
* et jour            adv
on                   spr,ag,sg,n;106
* est                v,sg,t
one                  nb
* man show           nn,m,sg
* step               nn,m,sg
orangs               nn,m,pl
* outans             nn,m,pl
* outangs            nn,m,pl
os                   ambi,nn,m,sg,nn,m,pl
* à moelle           nn,m,an
page                 nn,ag,sg
* un                 nn,f,sg
papiers              nn,m,pl
* émeri              nn,m,pl
paquet               nn,m,sg
* diy os             nn,m,sg
par                  prep;89,97,102
* dessous            prep;97,112
* dessus             prep;97,112
* devant             prep;97,112
* hasard             adv
* rapport à          prep;97,112
* rapport au         prep;97
* rapport aux        prep,z;97
paragraphe           nn,m,sg
* un                 nn,m,sg
parce                sub
* kiy                sub
* que                sub
parents              nn,m,pl
* et alliés          nn,m,pl
pas                  nnadv,z;83
* à pas              adv
pendant              ppadnant;97,102,112
* kiy                sub
* que                sub
père                 nn,m,sg
* et fils            nn,m,pl
petit                adj,m,sg
* à *                adv
* fils               nn,m,sg
peu                  adv
* de                 det
* diy                det
Ph.                  prnn
* de Champaigne      prnn
* Pétain             prnn
pick                 nn,m,sg
* up                 nn,m,an
pieds                nn,m,pl
* et poings liés     adv
pin                  nn,m,sg
* up                 nn,f,an
piper                inf
* cub                nn,m,sg
pis                  nnadv
* aller              nn,m,an
place                nnvb
* saint marc         nn,f,sg
plus                 ambi,adv,vbadvpp,z;83
* de                 det
* diy                det
* kiy                adv
* ou moins           adv
* que parfait        nn,m,sg
* que                adv
* tard               adv
* tôt                adv
* value              nn,f,sg
plût                 v,sg
* à dieu             adv
* au ciel            adv
plutôt               adv
* kiy                sub
* que                sub
poids                nn,m,an
* et haltères        nn,m,pl
* et mesures         nn,m,pl
pont                 nn,m,sg
* à Mousson          prnn
* Audemer            prnn
* aux ânes           nn,m,sg
* aux Dames          prnn
* Aven               prnn
* Euxin              prnn
ponts                nn,m,pl
* et chaussées       nn,m,pl
porc                 nn,m,sg
* épic               nn,m,sg
porcs                nn,m,pl
* épics              nn,m,pl
pot                  nn,m,sg
* à eau              nn,m,sg
```

* au feu	nn,m,sg
* au lait	nn,m,sg
* au noir	nn,m,sg
* aux roses	nn,m,sg
pots	nn,m,pl
* à eau	nn,m,pl
pour	prep;89,97,102
* kiy	sub
* que	sub
pourvu	ptp
* kiy	sub
* que	sub
près	adv
* de	prep;89,93,97
* des	prep,z;97
* diy	prep;89,93,97
* du	prep;97
prêt	adjnn,m,sg
* à porter	nn,m,sg
proche	nnadjadv
* de	prep;97,112
* des	prep,z;97
* diy	prep;97
* du	prep;97
profits	nn,m,pl
* et pertes	nn,m,pl
puis	coord,z
* je	v,sg
pull	nn,m,sg
* over	nn,m,sg
* overs	nn,m,pl
punching	adj
* ball	nn,m,sg
* balls	nn,m,pl
quand	sub,t
* est ce kiy	sub
* est ce que	sub
quant	prep,t
* à	prep;97
* au	prep;97
* aux	prep,z;97
quarante	nb
* huit	nb,tn
quatre	nb
* vingts	nb,z
* vingt un	nb,n
* * une	nb
* * deux	nb,z
* * trois	nb,z
* * quatre	nb
* * cinq	nb,tn
* * six	nb,tn
* * sept	nb
* * huit	nb,tn
* * neuf ans	nn,m,pl
* * neuf heures	nn,f,pl
* * neuf	nb
* * onze	nb
* * douze	nb
* * treize	nb
* * quatorze	nb
* * quinze	nb
* * seize	nb
* * dix sept	nb
* * * huit	nb,tn
* * * neuf ans	nn,m,pl
* * * neuf heures	nn,f,pl
* * dix neuf	nb
* * dix	nb,tn
quelque	adjadv;94,95
* chose	pr,sg
quelques	det,ag,pl,z
* unes	nn,f,pl;104
* uns	nn,m,pl;104
quitte	v,sg
* à	prep;89
R	nn,m,sg
* et D	nn,sg
rad	nn,m,sg
* soc	nn,m,sg
respect	nn,m,sg
* humain	nn,m,sg
rez	nn,m,an
* de chaussée	nn,m,an
ris	nnvb
* Orangis	prnn
sac	nn,m,sg
* diy os	nn,m,sg
saisies	nn,f,pl
* arrêts	nn,f,pl
Salt	prnn
* Lake City	prnn
San	prnn
* Diego	prnn
sang	nn,m,sg
* et eau	adv
sans	prep,z;89,97,102
* crit	nn,m,sg
* kiy	sub
* que	sub
sauf	prepadj;89,102
* kiy	sub
* que	sub
scène	nn,f,sg
* un	nn,f,sg
sciences	nn,f,pl

* nat	nn,f,pl
selon	prep;97,102,112
* kiy	sub
* que	sub
semen	nn,m,sg
* contra	nn,m,an
semper	adv
* virens	nn,m,an
sens	ambi,nn,m,an,v,sg
* dessus dessous	adv
* devant derrière	adv
serre	nnvb
* fils	nn,m,an
sex	nn,m,an
* appeal	nn,m,sg
shake	nn,m,sg
* hand	nn,m,an
show	nn,m,sg
* biz	nn,m,an
si	conjadv
* bien kiy	sub
* bien que	sub
sine	prep
* die	adv
* qua non	adv
sit	nn
* in	nn,m,sg
sitôt	adv
* kiy	sub
* que	sub
six	nb,tn
* quatre deux	nn,f,sg
siy	di,ag,an;84, 88,90,100
* est	v,sg,t
soixante	nb
* huit	nb,tn
statu	nn,m,sg
* quo	nn,m,sg
steeple	nn,m,sg
* chase	nn,m,sg
strip	nn,m,sg
* tease	nn,m,sg
strophe	nn,f,sg
* un	nn,f,sg
sud	nn,m,sg
* africain	nn,m,sg
* africaine	nn,f,sg
* africaines	nn,f,pl
* africains	nn,m,pl
* américain	nn,m,sg
* américaine	nn,f,sg
* américaines	nn,f,pl
* américains	nn,m,pl
* asiatique	adj,ag,sg
* asiatiques	adj,ag,pl
* est	nn,m,sg
* ouest	nn,m,sg
suivant	ppadnant
* kiy	sub
* que	sub
sus	vbadvpp
* dominante	adjnn,f,sg
* dominantes	adjnn,f,pl
* énoncé	adj,m,sg
* énoncée	adj,f,sg
* énoncées	adj,f,pl
* énoncés	adj,m,pl
* hépatique	adjnn,ag,sg
* hépatiques	adjnn,ag,pl
* maxillaire	adjnn,ag,sg
* maxillaires	adjnn,ag,pl
szx	prep
* chef	nn,m,sg
* lieutenant	nn,m,sg
* sol	nn,m,sg
t	nn,m,sg
* à	prep
* elle	spr,f,sg
* en	prep
* il	spr,m,sg
* on	spr,ag,sg
* shirt	nn,m,sg
tan	nn,m,sg
* sad	nn,m,sg
tandis	adv
* kiy	sub
* que	sub
tant	adv,t
* et plus	adv
* kiy	sub
* que	sub
Tchang	prnn
* Kaï Chek	prnn
te	di,sg;84,88,90,100
* deum	nn,m,sg
tee	nn,m,sg
* shirt	nn,m,sg
teen	nn,m,sg
* ager	nn,ag,sg
* agers	nn,ag,pl
tellement	adv
* kiy	sub
* que	sub
Th.	prnn
* de la Ville	nn,m,sg

```
* de Poche       nn,m,sg
* Français       nn,m,sg
tiers            adjnn,m,an
* arbitre        nn,m,sg
* état           nn,m,sg
* ordre          nn,m,sg
tissus           nn,m,pl
* éponges        nn,m,pl
tiy              di,ag,sg;79,84,88,
                 90,99,100
* est            v,sg,t
tome             nn,m,sg
* un             nn,m,sg
tôt              adv
* ou tard        adv
tous             ambi,pr,m,pl,det,
                 m,pl,z;94,96,104
* ces            det,m,pl,z
* égards         nn,m,pl
* les            det,m,pl,z
* leurs          det,m,pl,z
* mes            det,m,pl,z
* nos            det,m,pl,z
* ses            det,m,pl,z
* tes            det,m,pl,z
* vos            det,m,pl,z
tout             nnadndp,t;94,96,103
* à coup         adv
* à fait         adv
* à liy avenant  adv
* à liy heure    adv
* au long        adv
* au moins       adv
* au plus        adv
* au tout        nn,m,sg
* ce             det,m,sg
* cet            det,m,sg
* et pour tout   nn,m,sg
* le             det,m,sg
* leur           det,m,sg
* liy            det,m,sg
* mon            det,m,sg,n
* notre          det,m,sg
* ou rien        adv
* son            det,m,sg,n
* ton            det,m,sg,n
* votre          det,m,sg
toute            adndp,f,sg;
                 94,103,106
* cette          det,f,sg
* la             det,f,sg
* leur           det,f,sg
* liy            det,f,sg
* ma             det,f,sg
* notre          det,f,sg
* sa             det,f,sg
* ta             det,f,sg
* votre          det,f,sg
toutes           adndp,f,pl,z;
                 94,96,104
* ces            det,f,pl,z
* les            det,f,pl,z
* leurs          det,f,pl,z
* mes            det,f,pl,z
* nos            det,f,pl,z
* ses            det,f,pl,z
* tes            det,f,pl,z
* vos            det,f,pl,z
trachées         nn,f,pl
* artères        nn,f,pl
trente           nb
* huit           nb,tn
trop             adv,p
* de             det
* diy            det
un               det,m,sg,n;103,106
* fait exprès    nn,m,sg
us               nn,m,pl
* et coutumes    nn,m,pl
vade             v
* mecum          nn,m,sg
vents            nn,m,pl
* et marées      nn,m,pl
vice             nn,m,sg
* -président     nn,m,sg
vingt            nb,t
* et un          nb,n
* et une         nb
* deux           nb,z
* trois          nb,z
* quatre         nb
* cinq           nb,tn
* six            nb,tn
* sept           nb
* huit           nb,tn
* neuf ans       nn,m,pl
* neuf heures    nn,f,pl
* neuf           nb
vins             nn,m,pl
* et liqueurs    nn,m,pl
vis              ambi,nn,f,an,v,sg
* à vis de       prep;97,112
* à vis des      prep,z;97
* à vis diy      prep;97
* à vis du       prep;97
* à vis          adv
```

volley	nn,m,sg	* end	nn,m,sg
* ball	nn,m,sg	y	ambi,adv,nn,m,sg;
volume	nn,m,sg		79,85,88,100,101
* un	nn,m,sg	* as	v,sg
vous	allpron,ag,pl,z;79,	yé	nn,m,sg
	84,88,92,96,98,100	*-yé	adj,ag,an
* est	v,sg,t	z	nn,m,sg
walk	nn,m,sg	* à	prep
* over	nn,m,sg	* en	prep
week	nn,f,sg		

APPENDICES

APPENDIX I:	List of words which can be used to test the accuracy of a text-to-speech program	159
APPENDIX II:	List of sentences which can be used to test the accuracy of a text-to-speech program	160
APPENDIX III:	Algorithm for end-of-line hyphens	162
APPENDIX IV:	Abbreviations which are rewritten as complete words	174
APPENDIX V:	Treatment of hyphenated words (except end-of-line hyphens)	179
APPENDIX VI:	Borrowings, neologisms, and slang	180
APPENDIX VII:	Procedures for the assignment of linguistic features	182
APPENDIX VIII:	List of linguistic features	187
APPENDIX IX:	Root table	190
APPENDIX X:	Prefix table	191
APPENDIX XI:	Table of endings: modifications and assignment of features	193
APPENDIX XII:	Algorithm for initial /ˇ/	194
APPENDIX XIII:	Distribution of /a/ and /ɑ/	196
APPENDIX XIV:	The consonant /ɲ/	199
APPENDIX XV:	Transcription of oe and æ	199
APPENDIX XVI:	Algorithm for the unstable vowel	201
APPENDIX XVII:	Description of linguistic features 77 through 112	204
APPENDIX XVIII:	Figures	208

APPENDIX I: TEST WORDS

The following words present the types of difficulties which a text-to-speech program must solve to give a thoroughly accurate transcription:

abdomen	abritions	absorptions	abstentions
accroc	achetions	acquerra	acquiescent
acquisitions	action	adaptions	adhérent
admonestions	adolescent	affidavit	ajustions
allaitions	aloès	amer	ameutions
amputions	annotions	antialcoolique	appendicite
asexuées	assistions	assomptions	asthmatiques
aulx	automnes	auxquelles	ave
balbutient	baobab	bastion	benzine
bibliobus	bienheureuses	bioxyde	bisexué
boomerang	caennais	capétien	capharnaüm
celluloïd	ces	cet	chahut
chaldéen	champlever	chiendent	chienlit
chrétienté	chuintions	comptions	conceptions
consanguinité	consortium	constations	consultions
contrastions	contresens	corpus	corruptions
cosinus	couramment	courons	courrier
courriez	courroie	courrons	couscous
cul	curaçao	décoctions	décryptions
démentions	démocraties	démocratiquement	désaxer
descriptions	desdites	désennuient	désenragent
désensablage	désensibilisant	désexualiser	désherbant
désigner	dessous	destitutions	devient
dévient	diffamer	diligemment	discrétions
disproportions	dissertions	distinctions	dolmen
dysenterie	ébruitions	écartions	ecchymose
écourtions	édictions	égyptienne	empruntions
encoignure	enfer	enflamment	enherbent
enivrants	ennemi	enorgueillissent	énormément
enrhumons	entrelacs	envient	équerra
escient	eues	exaltions	exploitions
express	faisabilité	femmelette	fichtre
flirt	fonctions	football	fuchsia
fusil	gageure	gangs	gars
gâtions	gemment	gentilshommes	ghesha
gnomes	gong	goûtions	groseillier
gymkhana	h	habillions	habitations
haïtienne	héraldique	héraut	herpès
hétérosexuel	heurtions	hihan	humus
idioties	imbécillité	immiscent	inertie
inintelligent	initient	innocemment	insertions
interrègne	isthme	joaillier	kilohertz
laïussaient	lapsus	laser	lentement
lesquelles	libyen	lichen	looping
mach	maïeutique	mainmise	maïzena
mars	mayonnaise	menhir	mesdemoiselles
métalinguistique	mixtion	montgolfière	montions
mourons	mourrons	multisinistre	narguent

néanmoins	nerf	onzième	ours
outremer	oyez	pagaïe	panhellénique
paon	parasismiques	parasol	patientent
pays	pénicilline	péripétie	persil
pointions	polyester	porc	postions
pouls	précautions	prédictions	prénotions
présalaire	prétextions	productions	profundis
prudemment	psychose	quadrature	quadriennaux
radioisotopes	raid	rebab	reconquiers
rédemptions	regrettions	réinsertions	remmener
renne	rescousse	résipiscence	ressac
ressaie	ressent	ressuiez	ressuyez
révoltions	revolver	rhumons	rut
saccharinées	sanctions	sangsue	saoul
satiété	sautions	scient	sculptions
sélections	septième	serpillière	shakespearien
shampooing	sixième	solennellement	sphincter
steak	stop	strontium	stupéfactions
substantiels	sud	suies	suiez
superhypnose	superovarié	supportions	susdit
suspections	tapecul	te	teintions
terminus	thermos	toubib	tournevis
tram	tranquillité	transsaharien	triceps
trochaïque	tuyautées	ubiquité	unisexe
uruguayen	vissent	vraisemblable	zinc
zut			

APPENDIX II: TEST SENTENCES

The following sentences present the types of difficulties which a text-to-speech transcriber must solve in order to give a thoroughly accurate transcription. Since we wanted to include several difficulties in each sentence, the meaning of some test sentences may seem rather strange.

1. Ce sont les miens auxquels vous vous intéressez et non pas ceux que lui et moi inventions alors que nos ambitions se limitaient à ce que les as du SMIG considèrent aujourd'hui un fait accompli faute de ne pas habiter dans un HLM près de Stendhal.
2. Je suis bien trop fier pour m'aventurer à me fier à ce drôle de m'as-tu-vu si content de lui que tous content des histoires où le vice-président le mit knock-out au moyen-orient. Un véritable chef-d'oeuvre cohérent qui tient du miracle! Un grand espoir pour cet homme fort digne de nouvelles attentions solennelles.
3. Cachez-les-y chez elle sans que le grand oiseau noir sous-jacent qui les y avait cachés s'aperçoive qu'il a été négligent en ne respectant pas l'art. 93 du 12.10.1893 qui interdit de rester sur son quant-à-soi quand tout est mis sens dessus dessous.
4. Sait-on où Jules a placé l'annuaire de la SNCF? Quant à moi, il me convient de l'ignorer.
5. Les riches oisifs y président et le supporter hésite à le supporter sans arrière-pensée et sans électron-volt.
6. Aux anciens étudiants comme à ceux qui les suivront j'apporte le

salut fraternel de la SFIO et de ces quatre-vingt-six sections qui font pour nous tous tous les travaux six jours sur dix.

7. Si ton ancien ton en paraît moins agréable, je suppose que c'est la faute du qu'en-dira-t-on, n'est-ce pas?
8. Et sur ce il partit fort en rogne car nul au fort d'Eu n'avait eu l'idée de les y placer dans un magasin de vins et liqueurs.
9. Dites-moi à qui je peux donner ce morceau que les gourmets nomment le sot-l'y-laisse.
10. Ah, chanté-je, qui m'aime me suive jusqu'à en mourir!
11. Quiconque le hait ne hait que lui et toute sa famille.
12. Lui aussi sans doute le fera bien que je m'y oppose de toutes mes forces d'homme chérissant encore la liberté conquise par nos héros.
13. Nous avons deux cents employés: cent un négligent leur travail, quatre-vingt-un ferment la porte, mais nous avons un président excellent et intelligent qui devient vieux et ne sait même plus la valeur de 15^2.
14. Je ne lui ai jamais dit qu'au cas où sa voiture tomberait en panne il pourrait m'appeler de jour comme de nuit au 233-22-10.
15. Si vous m'aviez prévenu, j'aurais attendu patiemment et sans colère jusqu'à ce que la radio et la télévision confirment les résultats avant de crier victoire.
16. Neuf enfants ayant neuf ans chacun sont ici depuis avant-hier et doivent repartir aujourd'hui faire 12 km d'agit-prop.
17. Voulions-nous mentions A et B ou options D, E, F?
18. Chante-t-il juste ou faux? Qu'importe pourvu que sa chanson soit spirituelle!
19. Je leur ai dit ceci: 2 + 2 = 4 et 4 - 2 = 2 ici comme ailleurs.
20. Sans H^2O aucune vie n'est possible puisqu'il en faut aux corps.
21. Je sais qu'il est de Paris et qu'il y habite, mais y est-il toute l'année?
22. Que l'alphabet est difficile! Y est-il une voyelle ou une consonne?
23. Quand vous jouez de vos instruments, son ton et ton son excellent et me réjouissent.
24. Quand vous jouez de vos instruments, son ton insolent et ton son violent émergent et forment un corps excellent.
25. Emouvoir les foules n'est point à la portée de tous.
26. Il fait un bruit infernal, on en convient; mais tout de même, il faut bien qu'on s'amuse à toute heure quand on a son âge et qu'on est plein de vie! D'ailleurs je vous assure que toute cette activité lui fera sans doute un bien fou.
27. Le vieux chef de gare et le garde-barrière égrotant qui nous accompagnent s'impatientent.
28. Arrivée à Paris au printemps de 1896, elle devint la reine de tous les bals mondains et ce grâce à son éclatante beauté et à son je-ne-sais-quoi.
29. Donne-leur le la si la si nouvelle harmonie te plaît plus qu'à moi.
30. Il sera un saint, mais en aura l'aura et le halo bien plus tard.
31. Les fils de ce président excellent et content négligent les inventions.
32. Ce quartier de la banlieue est de Lyon est depuis toujours un foyer d'agressions diverses.
33. Regardez le point M. Dubois nous a dit que le boul. Marat, qui a 25 m. de large, est l'artère préférée de MM. Ch. Lidoy et Ed. Ibert.
34. Après deux années de Normale Sup', il quitta Paris le 28.5.1980 pour

s'installer 13, Av. des Pyrénées et consulter l'édit. de 1835.
35. Rodez votre voiture à Rodez. C'est bien mieux que sur l'autoroute A6; il y a plus de côtes au km^2. C'est M. de B***, ce grand <<idéologue>>, qui me l'a conseillé pour la 1re fois.
36. Dans le chapitre I, la S.F.I.O. a dépensé 978,45 F T.T.C. À 12:15 h.
37. Achetez-en ici, allez en chercher à Brest, osez y aller, envoyez-les acheter du pain, et regardez-nous écrire.
38. Le vice-président dit que le crime et le vice président à leurs actes.
39. Si mon frère, ma soeur, mes fils, tous mes amis et nos collègues n'avaient jamais rien su de ce que fut ma vie avant que ma famille ne vînt s'établir dans ce pays béni des dieux tout de suite après le cruel holocauste qui rappelle les atrocités de Louis X et de Louis XIV, toutes les autorités morales de ce pays m'auraient refusé ce sans quoi je ne puis vivre; debout, j'aurais alors pris fait et cause pour ces vampires d'outre-tombe à l'UNESCO.
40. Il montra une broche et nous dit qu'il l'avait soustraite à un certain bijoutier.
41. Après avoir signé les contrats, il admit qu'il soustraite la fabrication des éléments électroniques.

APPENDIX III: ALGORITHM FOR END-OF-LINE HYPHENS

A. Hyphens that occur at the end of a line are of two different types. Some would disappear if the two strings that are separated by the end-of-line hyphen were on the same line (e.g., mon-sieur, rem-place) and some would remain (e.g., dix-sept, ceux-là). We will use the term p-hyphen (permanent hyphen) to designate the hyphens that would remain and the term t-hyphen (temporary hyphen) to designate the hyphens that would disappear. The potential problems presented by end-of-line hyphens (assignment of erroneous linguistic features and wrong pronunciations) are illustrated by these four pairs of sentences (in the first sentence of each pair, we have a t-hyphen; in the second one, we have a p-hyphen):

Il nous montra une broche et nous dit qu'il l'avait soustraite à un certain bijoutier ...
Après avoir signé tous les contrats, il admit qu'il soustraite la fabrication des éléments ...

Leur offensive avait commencé dès l'aube; il fallait culbuter l'ennemi ...
Rien ne bougeait dans cette sombre clairière sauf un culblanc qui construisait son nid .....

Il voulait bien accepter des réformes, mais pas la chienlit ...
Il les adorait et il avait toujours près de lui un chienloup ...

Elle a toujours eu de bonnes notes en classe sauf en calcul ...
Il était plutôt vulgaire et la traitait souvent de casse-cul ...

How serious are those problems? For example, what effect would it have on the accuracy level of our program if we simply removed all end-of-line hyphens and joined the two strings into one? In order to measure this percentage of errors, we chose some books (texts 1-4) and some news articles (texts 5-7) at random:

1. Opéra Conjugal by José-André Lacour, La Table Ronde, Paris, 1978
2. Une Oreille dans le Train by Robert Cransac, Denoël, Paris, 1966
3. Histoire des Indiens d'Amérique du Nord by Philippe Jacquin, Payot, Paris, 1976
4. Cathédrales de France by Pierre Gérard, Nathan, Paris, 1962
5. Midi Libre (July 13, 1983)
6. Le Canard Enchaîné (January 25, 1984)
7. L'Express (May 27 - June 2, 1983)

In those texts, we found a total of 89 p-hyphens (when the example has more than one hyphen, we use = to indicate the end-of-line hyphen):

Text 1: demi-effacées, chef-d'oeuvre, peut-être, salle-à=manger, avoua-t=elle, grand-mère, Saint-Brévin, Saint-Pierre, rendez-vous, Saint-Pierre=La-Barre, demi-heure

Text 2: quatre-vingt, Notre-Dame, savez-vous, dimanche-là, êtes-vous, voulez-vous, elle-même, croyez-vous, abat-jour, queue=de-poisson, Opéra-Comique, dix-neuf, veux-tu, demi-tour, beau-père, extra-doux

Text 3: Etats-Unis, Nord-Ouest, 1595-1617, Oiseau-Tonnerre, Sud-Ouest, homme-médecine

Text 4: Vaison-la=Romaine, Marie-Madeleine, arcs-boutants, rendez-vous, bas-reliefs (twice), Notre-Dame (five times), 1155-1174, elle-même, Jean-Baptiste (twice), Sainte-Ampoule, Saint-Guy, Saint-Etienne, Saint-Jean, Ile-de=France (twice), Saint-Sernin, vice-gérant, Prince-Noir, Saint-Denis, Clermont-Lodève, Viollet-le=Duc (twice)

Text 5: Etats-Unis, Estaque-Gare, Radio-Tour, Fontaine-au=Pire

Text 6: quelques-uns, semi-défait, Saint-Chamond, Etats-Unis, FR3-Corse, fac-similé, demi-douzaine, Nord=Pas-de-Calais, peut-être

Text 7: avant-garde, européo-américain, porte-parole, Citroën-Aulnay, lui-même (twice), Jean-François, anglo-américain, sont-ils, auraient-ils, Alpes-Côtes d'Azur, anti-impérialistes, Honeywell-Bull, politico-sociale, au-dessus

The distribution of the two types of end-of-line hyphens is as follows:

Text	Total number of hyphens	Number of t-hyphens	Number of p-hyphens	Percentage of p-hyphens
1	90	79	11	12.22%
2	465	449	16	3.44%
3	138	132	6	4.34%
4	350	322	28	8.00%
5	213	209	4	1.87%
6	282	273	9	3.19%
7	360	345	15	4.16%
	1898	1809	89	4.68%

We then measured the ratio of end-of-line hyphens to phonemes. In newspapers or magazines with narrow columns, we found about one end-of-line hyphen per 100 phonemes; in books, the ratio decreased to about one end-of-line hyphen per 550 phonemes.

Since removing all the end-of-line hyphens and joining the two strings into one ensures that all the words which were divided by a t-hyphen will be analyzed and transcribed correctly, the errors can occur only with the 89 p-hyphens. Removing these hyphens does create havoc with the linguistic analysis (e.g., dimanchelà, rendezvous), but —if we consider only the pronunciation— we find that our computer program yields a correct transcription for 51 items.

Thus, under the worst conditions (one end-of-line hyphen per 100 phonemes), the error rate would be 1/100 x 4.68/100 x 38/89 = 2 errors in 10,000 phonemes; under the best conditions (one end-of-line hyphen per 550 phonemes), the error rate would be 1/550 x 4.68/100 x 38/89 = 0.36 error per 10,000 phonemes. In terms of reading a text aloud (average speed of ten phonemes per second), the elimination of all end-of-line hyphens would result in one audible error every eight minutes (under the worst conditions) or one audible error every forty-five minutes (under the best conditions).

Although this error rate is extremely small, we must try to eliminate it.

B. In order to find a solution, we considered these two possibilities:

1. One approach is of a semantic nature. Its single rule postulates that end-of-line hyphens which occur between lexical items (e.g., soutien-gorge) are p-hyphens and that all others are t-hyphens. The list of exceptions would include some p-hyphens that would not be caught by the rule (e.g., ping-pong, tam-tam, tom-pouce) and some t-hyphens that the rule would interpret as p-hyphens (e.g., long-temps, cou-vent, pomme-raie). The list of exceptions would have to be established with great care since failing to place some items on the list would result in wrong feature assignment (e.g., garde-rie, panse-ment, mine-rai, long-temps, pomme-raie) and, more seriously, in wrong pronunciations (e.g., cal-cul, con-ter, cou-vent, dis-tance, en-tiers, jus-que, para-des, si-nus, sous-traction, ta-chine, ta-maris, un-ciné, ver-sus).

This approach encounters two major obstacles:

a. It requires an enormous data base (see page 19) containing all the forms of all the words used in French.

b. While it is true that the vast majority of p-hyphens occur between lexical items (e.g., timbre-poste), it is not true that most hyphens which occur between strings which can be lexical items are p-hyphens (e.g., long-temps, cou-vent). The list of exceptions would contain thousands of items; for example, the Petit Robert (1979) has 209 entries beginning with dis and in 82 cases the string following dis can be a lexical item (e.g., dis-corde, dis-cours, dis-section). Even if we wanted to list only the exceptions which present a pronunciation problem (e.g., dis-continuer), we still would face a formidable task since we would have

to determine whether each of the entries in the data base can be the initial syllable(s) (e.g., en, auto) of another entry in the data base (e.g., entiers, automobile), whether the non-initial syllables (e.g., tiers, mobile) are separate entries in the data base, and whether the pronunciation of the whole string is the same (e.g., automobile) or not the same (e.g., entiers) as the sum of the pronunciations of the separate strings (e.g., auto + mobile, en + tiers).

2. Another approach considers the letters which precede and follow the hyphen and attempts to determine which combinations have t-hyphens and which ones have p-hyphens. From a linguistic point of view, this approach appears more sound since there is a finite set of letter combinations. The number of combinations is very high since we have to consider vowel-vowel, vowel-consonant, consonant-vowel, three consonants-consonant(s), two consonants-consonant(s), one consonant-consonant(s), but this should not be a problem. The difficulty we have found is that many combinations occur with t-hyphens as well as with p-hyphens. This is particularly true with the sequence one consonant-consonant(s). Examples:

c-c	ac-cident, oc-cuper	bec-croisé, bloc-cuisine
c-s	toc-sin	fac-similé
c-t	ac-tiver	tic-tac
d-c	God-chaux	sud-coréen
d-n	échid-né	pied-noir
f-t	chef-taine	teuf-teuf
l-b	cul-bute	cul-blanc
l-d	sol-dat	social-démocrate
l-m	cal-mer	fusil-mitrailleur
l-v	hel-vête, sal-vateur	col-vert, cheval-vapeur
m-m	fem-mes	miam-miam
m-p	cam-pagne, rem-place	tom-pouce
m-t	com-té	tam-tam
n-b	bon-bonne	cordon-bleu
n-c	nan-céen, bron-chite	semen-contra, camion-citerne
n-d	on-dée, in-demne	on-dit, bien-dire, in-douze
n-f	sain-foin, in-fernal	jean-foutre
n-g	lin-got	soutien-gorge
n-h	bon-heur, Pan-hard	Ben-Hur
n-l	en-lève, chien-lit	chien-loup
n-m	néan-moins, main-mise, Pen-march	maison-mère, bain-marie, colin-maillard
n-n	mon-naie, in-né	écrivain-né
n-q	cin-quième	in-quarto
n-r	vin-rent	non-recevoir
n-s	mon-sieur, en-semble	non-sens, poisson-scie, in-seize
n-t	en-têté, gen-tilhomme	en-tête
p-p	ap-plaudir, sup-porter	trop-plein, trop-perçu
p-t	sep-tième	strip-tease
r-b	Ser-bie, ar-brisseau	fer-blanc
r-c	par-courir, mer-ci, cher-che, démar-cheur	tiroir-caisse, piper-cub, par-ci, laurier-cerise, ver-coquin
r-d	char-don	dîner-débat
r-f	cer-feuil	papier-filtre

r-l	per-le	par-là
r-m	per-mettre, désor-mais	quartier-maître, papier-monnaie
r-n	der-nier	dernier-né, premier-né
r-p	pour-point, ar-pent	trésorier-payeur, amour-propre
r-r	car-refour	laurier-rose
r-s	per-san, jer-sey	pur-sang
r-t	par-tir	laurier-tin
r-v	par-venir, conser-vent	savoir-vivre
s-b	pres-byte, gris-bi, as-bestose, es-broufe, les-bienne, Gros-bois, Lis-bonne	paris-brest, gris-bleu, bas-bleu, sabres-baïonnettes, bouis-bouis, tapis-brosse, appuis-bras, moissonneuses-batteuses, gros-bec
s-c	res-capé, manus-crit, adoles-cent, bis-cuit, dis-cerner, sus-ceptible	lettres-consonnes, bas-côté, basses-cours, aides-comptables, reines-claudes, agnus-castus
s-d	des-dits, mes-dames, les-dites	villes-dortoir, gras-double, agnus-dei, aigres-douces, après-demain
s-g	vos-gien	gros-grain, après-guerre
s-h	dés-habillé	ampères-heures, trois-huit
s-j	dis-joint, Des-jardins	bas-jointé(e)(s), dames-jeannes
s-l	is-lamique, Stanis-las, Ghis-laine	chênes-liège, années-lumière, compères-loriots, cartes-lettres
s-m	os-mose, bis-muth, Huys-mans, cos-modrome, charis-matique	villes-musée, bas-mât, sous-main, sus-maxillaire, trois-mâts, après-midi, trous-madame, cuisses-madame
s-p	as-phalte, es-poir, res-pect, dis-paraître	louis-philippard, timbres-poste, pelles-pioches, surprises-parties
s-q	des-quels, jus-que	louis-quatorzien
s-r	is-raélien, Des-rosiers	bas-relief, poches-revolver
s-s	gros-sir, lais-sent, mes-sieurs, res-source,	chauves-souris, hommes-sandwiches, pochettes-surprises, trois-six, après-ski
s-t	res-ter, bas-tion	basses-tailles
s-v	Des-vallières, Bois-vin	bas-ventre, procès-verbal, plus-value, assurances-vieillesse
t-b	haut-bois	basket-ball, petit-beurre
t-c	pat-chouli	lit-cage, mot-clé, avocat-conseil
t-g	cat-gut	petit-gris
t-m	at-mosphère, leit-motiv	état-major, petit-maître, bracelet-montre
t-n	Et-na	petit-neveu, petit-nègre
t-p	Bet-plan	haut-parleur, pur fruit-pur sucre
t-s	Bat-sère	petit-suisse, t-shirt
x-c	ex-cès	chapeaux-cloches
x-p	ex-pression	ex-professeur, faux-pas, beaux-pères
x-s	ex-sangue	faux-semblant, dix-sept
x-t	jux-taposer, sex-tuple, immix-tion, bissex-tile	fox-trot, veux-tu, dix-tonnes, moteur deux-temps

The above examples show that it is not possible to write rules which would be based only on letter combinations. The rules we have put together use word lists and letter combinations. They probably do not solve all the problems posed by end-of-line hyphens, but our rules are correct in clearly over 99% of the cases, which means that the maximum number of errors due to end-of-line hyphens is two in one million phonemes.

C. A minor difficulty in writing the rules is due to the fact that not all printers agree on the position of some syllabic cuts; for example, we find cons-tant, subs-tantif, manus-crit, sans-crit as well as con-stant, sub-stantif, manu-scrit, san-scrit. [See Grevisse (1980), section 111.] Our rules give the correct transcription for both types of syllabic cut and all such words appear under two rules.

The rules are to be applied in the order given below. When an example contains several hyphens, we use = for the hyphen(s) concerned by the rule. [At this point in our program, several rewritings have already taken place; for clarity, the following rules use the unmodified spellings.]

RULE 1: If the string preceding the hyphen is space + a prefix from our table of prefixes, we remove the hyphen and we join the two strings into one. Some hyphens are p-hyphens (e.g., néo-impressionisme) and some are t-hyphens (e.g., parasol); in either case, the phonemic transcription and the linguistic features are correct.

con-struction	inter-stice	néo-impressionisme	re-cul
trans-humance	super-étendard	oxy-acétylénique	con-stant
radio-isotope	post-poser	micro-organisme	para-sol
électro-aimant	auto-école	psycho-social	

The same restrictions as those shown in the prefix table (pages 191-192) apply (for example, entre-nerf is treated on page 171).

RULE 2: This rule catches the most frequent occurrences of p-hyphens. It catches also a very few t-hyphens, but in each case the pronunciation is correct and there is no significant error in the assignment of features.

The hyphen is replaced with a space and the features "prehyph" and "posthyph" are set when:

a. The letter preceding the hyphen is a consonant and the letter following the hyphen is a vowel. In this case, only p-hyphens occur:

saint-esprit	bien-aimé	dit-elle	prends-en
chantent-ils	vend-on	moyen-âge	tout=à-l'égout
allez-y	vous-en	les-y	non-activité
Etats-Unis	tiers-état	Michel-Ange	suisses-allemands
guet-apens	en-avant	laisser-aller	quatre-vingt=un
porc-épic	sans-abri	vif-argent	carton-amiante
quelques-uns	pied=à-terre	mot-outil	machines-outils
arc=en-ciel	trois-étoiles		

b. The hyphen is part of one of the following strings (all are p-hyphens except in aujour-d'hui):

-t-	dira=t-on, aura=t-elle, chante=t-il
-à-	prêt-à=porter, bourses-à=pasteur, vis-à=vis, faces-à=main, acquit-à=caution, quant-à=soi, tête=à=tête, touche=à=tout tête=à=queue, corps-à=corps
-au-	pot-au=feu, vol-au=vent, tire=au=cul

-aux-	mort-aux-rats
-de-	dessous-de-plat, dessus-de-lit, lie-de-vin, esprit-de-vin, pas-de-porte, cul-de-sac, oeils-de-boeuf, pet-de-nonne
-d'	chef-d'oeuvre, dos-d'âne, pas-d'âne, aujour-d'hui [Splitting aujourd'hui into three words (aujour, diy, hui) does not interfere with the transcription.]
-du-	monnaie-du-pape, bonheur-du-jour
-des-	reine-des-prés, trompette-des-morts
-en-	croc-en-jambe, arc-en-ciel, pet-en-l'air
-et-	va-et-vient
-le-	sans-le-sou, cessez-le-feu, bras-le-corps, sur-le-champ, Louis-le-Grand, dis-le-moi, chante-le-nous, haut-le-corps
-la-	trompe-la-mort, trompette-de-la-mort, hors-la-loi, prête-la-lui, rappelle-la-toi
-les-	rappelez-les-vous, prêtez-les-leur
-l'	pont-l'évêque, trompe-l'oeil, sot-l'y-laisse
-t'	sers-t'en
-m'	prête-m'en, je-m'en-fichisme
-que-	plus-que-parfait
-z-	quatre-z-yeux

c. The hyphen occurs before a capital letter. All are p-hyphens:

Charles-Martel Jean-Pierre Ben-Hur rue Victor-Hugo

d. The hyphen occurs between:

-- a non-accented e and a vowel. In this case, the hyphens are p-hyphens since divisions such as mange-âtes, se-izième, mange-ons, de-ux are not acceptable. Examples:

contre-espionnage	cure-oreille	machine-outil	vice-amiral
extrême-oriental	morte-eau	pèse-acide	vide-ordures
porte-avions	maître-autel	libre-échange	porte-épée

-- a and i, a and u, i and a, i and e (accented or not), i and i, i and o, i and u, i and y, o and e, o and i, o and u, o and y:

bi-acide	génito-urinaire	ci-après	bi-univoque
demi-effacé	franco-yougoslave	intra-utérin	milli-ampère
demi-étape	gastro-intestinal	iso-ionique	indo-européen
jéjuno-iléon	parlez-lui-en	péri-urbain	sacro-iliaque

[Only a few examples are t-hyphens (e.g., bi-acide, péri-urbain).]

e. The hyphen follows or precedes a numeric character. Examples:

Pacific-231 Pr3-Normandie 456-179 Europe-1

f. The hyphen occurs between:

-- The letters e, s, or z and any of the words moi, toi, le, la,

les, lui, leur, nous, vous, soi. Examples:

dis-moi parle-leur taisez-vous rendez-vous chez-soi

-- The letters s, x, or z and the words tu, nous, or vous. Examples:

as-tu dis-nous faites-vous servez-vous veux-tu

-- The letter t and the word ce. Examples:

sont-ce était-ce n'est-ce pas

-- The letter c or p and the letter h. Example:

cap-hornier

g. The hyphen follows à. Examples:

là-bas là-dedans à-coup déjà-vu

h. The hyphen follows the letters aux, eux, oux. Examples:

faux-pas feux-follets choux-raves

i. The hyphen occurs after any single consonant except "l, n, r, s" and before l or r (p-hyphens since this division is not acceptable in a word):

mandat-lettre	décret-loi	chef-lieu	huit-reflets
sub-lingual	bout-rimé	deux-roues	croix-rouge
petit-lait	new-look		

RULE 3: The hyphen is replaced with a space and the features "prehyph" and "posthyph" are set when the hyphen is preceded by one of these words:

(Some words are to be found only when the word following the hyphen does not begin with a particular letter since, obviously, we do not want to keep mot and tes, fusil and lade, dix and mude, etc. separated. This letter is shown in parentheses.)

abat(t)	agnus	aide	aides	aigre
aigres	américano	ampère	ampères	anarcho
anglo	année	années	apprenti	apprentis(s)
appui	appuie	appuis	après	arrière
artiste	artistes	assurance	assurances	autos
avant	avocat	âme	âmes	bande
bandes	basket	basses(s)	basse(m)	bat(t)
belles	belle(m)	billet(t)	blanc	bloc
boui	bouis	bout	boute	bracelet
brise(m)	cache	caches	carte	cartes
casque	casques	casse(m)	chasseur	chat(t)
chaud	chauve	chauves	châssis	chef(f)
chêne	chênes	chirurgien(n)	ci	cinquante
cité	cités	coffre	coffres	comédie

comédies	commis(s)	compère	compères	contre(s)
cortico	cuisse	cuisses	dame	dames
demi	dernier	deux	disque	disques
dix(m)	dîner	droit	enquête	européo
écrivain	état	fait	femme	femmes
fille	filles	fox	franc	franco
frère	frères	friction	fruit	fusée
fusées	fusilier	fusil(l)	garde	gardes
gardien(n)	gâte	germano	globe	grand
gras(s)	gratte(m)	grippe	gris(h)	gros(s)
haut	hispano	homme	hommes	horloger
hors	hôtel	hôtelier	ici	idée
idées	image	images	italo	jupe
jupes	kilomètre	kilomètres	laissez	lampe
lampes	langue	langues	laurier	lettre
lettres	lèche(m)	libres	libre(m)	lit(t)
livre	livres	louis	loup	marie
mezza	mezzo	mécanicien(n)	médico	médullo
moine	moines	moins	moissonneuse	moissonneuses
monte	mot(t)	music	non(c)(n)	officier
oiseau	one	ouvrier	palais	panier
papier	paquebot	pare(m)	paris	pays
pelle	pelles	petit	petite	petites(s)
pèse	pied	pince	pinces	pique
pistolet	plat	plate	plates	plus
poche	poches	pochette	pochettes	poisson(n)
politico	porte	portes	pot(t)(l)	premier
prix	procès	quarante	quartier	quatre
rabat(t)	reine	reines	repas	réveil
riz	rois	russo	sabre	sabres
sacro	saint	salaire	salaires	sans
sellier	semi	serviette	serviettes	show
soeur	soixante	steeple	strip	surprise
surprises	tabac	tape(m)	tapis(s)	ticket
timbre	timbres	tire	torche	tout
trente	trésorier	trois	trou	trous(s)
ultra	vélo	vélos	village	villages
ville	villes	voiture	voitures	volley
voyage	voyages	water		

Some words are erroneously caught by this rule (e.g., aide-rait, ci-té). In all cases, the phonemic transcription is correct; the erroneous linguistic assignments do not have a significant effect on the division of the input string into breath-groups or on the liaisons.

Since some writers and printers separate sanscrit as sans-crit as well as san-scrit, we have placed sans-crit on the idiom list of the lexicon.

RULE 4: The hyphen is replaced with a space and the features "prehyph" and "posthyph" are set when the hyphen is followed by one of these words:

blanc	blancs	boeuf	boeufs	chef
chefs	ci	ciment	ciments	cinq

club	clubs	contre	corps	côté
côtés	croisé	croisée	croisées	croisés
cub	cubs	cumulus	dehors	dernier
derniers	dessous	dessus	deux	dix
douze	femme	femmes	fond	fonds
gorge	gorges	grand	grands	hépatique
hépatiques	homme	hommes	huit	jacent
jacents	je	joaillier	joailliers	jointé
jointée	jointées	jointés	jong	là
limite	limites	lit	lits	loup
loups	madame	madames	mât	mâts
mecum	mesdames	messieurs	même	mêmes
miam	monsieur	monsieurs	nerf	nerfs
neuf	nimbus	onze	pace	pied
pieds	pilote	pilotes	plano	pouce
pouces	premier	premiers	président	présidents
quarto	quatorze	quatre	quinze	relief
reliefs	revolver	revolvers	sanguin	sanguins
seize	sept	sébacé	sébacée	sébacées
sébacés	shirt	shirts	six	soldat
soldats	soleil	souvenir	souvenirs	soviétique
soviétiques	stratus	suicide	suicides	support
supports	surprise	surprises	tabac	tam
tams	tasse	tasseuse	tasseuses	témoin
témoins	treize	trente	trois	type
types	un	vapeur	vent	ventre
ventres	vents	vingt	vingts	

Some words are erroneously caught by this rule (e.g., mer-ci, adou-ci, voi-ci, voi-là, hi-deux, oc-trois, Ca-dix, ra-pace, cara-pace, canta-loup pla-fond, tré-fonds); the pronunciation is not affected except for Cadix which has been placed on the list of idioms (under Ca). The erroneous linguistic assignments do not have a significant effect on the division of the input string into breath-groups or on the liaisons.

RULE 5: In all remaining occurrences of vowel + hyphen + vowel, we remove the hyphen and we join the two strings into one:

nancé-en agré-able yé-yé sino-américain afro-asiatique

The transcription and the linguistic analysis are always correct for t-hyphens. All the examples of p-hyphens which we found were transcribed correctly and there were no significant errors in the linguistic analysis. There is only one potential problem; should the string following the p-hyphen be in the lexicon because our transcription rules would not transcribe it correctly, the dehyphenated combination would be mispronounced.

RULE 6: This rule deals with the remaining cases of vowel + hyphen + consonant(s). We remove the hyphen and we join the two strings into one. If the hyphen is a t-hyphen, the resulting dehyphenated combination is always transcribed and analyzed correctly:

pari-sien basse-ment ca-hot casse-ment

certi-fier	fai-sons	gêné-rations	manu-scrit
occa-sion	bi-seau	bi-sexué	déca-syllabique

Except for à-peu-près, the combinations with p-hyphens are transcribed correctly and the linguistic analysis is correct in most cases:

amuse-gueule	gagne-pain	au-deça	au-dedans
au-delà	au-devant	bateau-mouche	beau-père
chauffe-plats	compte-tours	mille-pattes	pêle-mêle
pause-café	quatorze-juillet		

To our knowledge, à-peu-près is the only hyphenated expression which, in this category, requires special treatment; it had to be added to Group 3 of the lexicon (as peuprès). Peu could not be entered under Rule 3 because of words like peuple and près could not be entered under Rule 4 because of exprès.

RULE 7: This rule deals with the remaining cases of three consonants + hyphen + consonant(s). We have not found any t-hyphens in this sequence (t-hyphens occur after the first or second consonant: ob-struer, feld-spath). Therefore, we replace all hyphens occurring after three consonants with a space and we set the features "prehyph" and "posthyph". Examples:

arcs-boutants	blancs-seings	cerfs-volants	corps-morts
francs-tireurs	grands-pères	ronds-points	sourds-muets
temps-mort	match-retour		

RULE 8: This rule deals with the remaining cases of two consonants + hyphen + consonant(s). Except for some proper names (e.g., Alfortville) and a very small number of words (e.g., arctique, longtemps, sangsue), we were able to separate the t-hyphens from the p-hyphens. In all cases, the transcription is correct and the erroneous linguistic assignments made for longtemps, sangsue, etc. do not have a significant effect on the division of the input string into breath-groups or on the liaisons.

a. Hyphen occurring after r + consonant: In this case, we replace the hyphen with a space and we set the features "prehyph" and "posthyph" (the combinations with hors are treated by Rule 3):

Alfort-ville	arc-boutant	arc-tique	Bourg-barré
cerf-volant	court-vêtu	court-bouillon	discours-programme
dîners-débats	vert-pomme	lauriers-roses	lauriers-cerises
messieurs-dames	mort-né	nord-coréen	nord-sud
quartiers-maîtres	port-salut	sourd-muet	tiers-monde
tiroirs-caisses	tord-boyaux	tord-nez	derniers-nés
premiers-nés			

Exceptions: We remove the hyphen and we join the two strings into one when the hyphen occurs in rs-p, rs-q, rg-m, and rt-s:

pers-pective	pers-picace	lors-que
bourg-mestre	hert-zien	

b. Hyphen occurring after n + consonant: In this case, we replace the hyphen with a space and we set the features "prehyph" and "posthyph" (the combinations with avant, blanc, franc, grand, moins, saint, sans are treated by Rule 3). Examples:

bains-marie	cent-garde	boutons-pression	chewing-gum
cent-suisse	Font-vieille	lieutenant-colonel	long-courrier
Long-jumeau	long-temps	Pont-chartrain	Mont-pellier
Mont-rouge	ping-pong	romans-feuilletons	point-virgule
pont-levis	punching-ball	stations-service	rond-point
sang-froid	sang-sue	wagons-restaurant	sent-bon

The combination bons-hommes will be transcribed correctly by our program since the liaison between bons and hommes will be made.

Exceptions: We remove the hyphen and we join the two strings into one when the hyphen occurs in nc-t, ng-t, nd-c, ns-c, ns-p, ns-t:

tinc-torial	mand-chou	cons-crit	ving-tième
cons-piration	ins-tant	cons-tater	

Succinctement is in the lexicon since it is pronounced /syksɛ̃tmɑ̃/.

c. Hyphen occurring after consonant (except r and n treated in previous rules) + s: In this case, we replace the hyphen with a space and we set the features "prehyph" and "posthyph":

avocats-conseils	billets-doux	blocs-cuisines
chefs-lieux	loups-garous	chats-huants
décrets-lois	états-majors	faits-divers
fusils-mitrailleurs	lits-cages	loups-cerviers
mandats-lettres	mots-clés	petits-suisses
pieds-noirs	poids-lourd	petits-beurre
culs-terreux	chauds-froids	bouts-rimés

The combination gentils-hommes will be transcribed correctly by our program since the liaison between gentils and hommes will be made.

Exceptions: We remove the hyphen and we join the two strings into one when the hyphen occurs in bs-t:

obs-tiné	subs-tantif	abs-traction

d. In the remaining cases, we remove the hyphen and we join the two strings into one. Examples:

ath-lète	comp-table	domp-teur	ryth-mique	sculp-teur
somp-tueux	exemp-ter	feld-spath	promp-titude	pick-pocket

RULE 9: This final rule deals with the remaining cases of one consonant + hyphen + consonant(s).

a. If the string preceding the hyphen is space + sous and if the string

following the hyphen is cri, sign, trac, trair, trais, trait + space, traits, traite + space, traites, or tray we remove the hyphen and we join the two strings into one. If the resulting string is soustraite or soustraites, we have to determine whether we have the past participle of soustraire or the present of sous-traiter; these two strings have been added to the list of items which must go through the disambiguation process. In all other occurrrences of space + sous + hyphen, we replace the hyphen with a space and we set the features "prehyph" and "posthyph":

sous-traiter sous-verre sous-marin sous-sol sous-titre

b. In the remaining cases, we remove the hyphen and we join the two strings into one. If the hyphen is a t-hyphen, the resulting dehyphenated combination is always transcribed and analyzed correctly:

mon-sieur dis-tance par-venir sep-tième

Most combinations with p-hyphens are caught by Rule 3 or by Rule 4. Except for bas-bleu, the remaining combinations are transcribed correctly and the erroneous linguistic assignments do not have a significant effect on the division of the input string into breath-groups or on the liaisons:

bec-fin	bien-dire	bon-point	bouton-pression
ex-ministre	main-forte	martin-pêcheur	on-dit
rad-soc	sus-maxillaire	tiroir-caisse	trop-plein

Bas-bleu has been added to Group 3 of the lexicon (as basbleu). Bas could not be entered under Rule 3 because of words like bastion and bleu could not be entered under Rule 4 because of palsambleu.

APPENDIX IV: ABBREVIATIONS REWRITTEN AS COMPLETE WORDS

We list only once the abbreviations in which the final letter(s) can appear either at line level or in superscript position (e.g., Mme or M^{me}).

Whenever possible, the spelling is simplified (e.g., boulvar) in order to speed up the transcription and/or provide more accurate features.

It is assumed that the following abbreviations are either initial of string or preceded by a space since we do not want the abbreviation M. to be found when O.N.M. is used.

The abbreviations which are listed without a period (e.g., hz, vx, mV) are those which we rewrite whether they appear with or without a period in the text; the abbreviations which are listed with a period (e.g., art., cal.) are those which we rewrite only if they appear with a period in the text. The third column lists the conditions under which the abbreviations are rewritten: pw. means preceding word; fw. means following word; "pok" means "period not a true punctuation mark" as is the case in:

... il a acheté l'édit. de 1978.
... votre lettre du 12 cour.;

a	are	if bit 74 = 1 in pw.
A	ampère	if bit 74 = 1 in pw.
AE, A.E.	afèrzétranjère	if pw. is les, aux, or des
AF, A.F.	anciinfran	if bit 74 = 1 in pw.
amp.	ampère	
appt	appartement	
arr.	arondisman	
art., Art.	article	if bit 74 = 1 in fw. or if "pok"
a/s	aubonsoinde	
atm.	atmosphère	
av.	avant	if fw. is J.-C. or J-C
av.	avnu	in other cases
Av.	avnu	
bcbg, BCBG	bonchicbonjanr	
bd, Bd	boulvar	
bld, Bld	boulvar	
B^{on}	baron	only with superscript on
B^{onne}	barone	only with superscript onne
boul., Boul.	boulvar	
BPF	bonpourfran	if bit 74 = 1 in fw.
B.P.F.	bonpourfran	if bit 74 = 1 in fw.
br.	broché	
B$_{e}$-du-R.	boushdurône	
B.-Rh.	bârin	
c	centime	if bit 74 = 1 in pw.
C	celsius	if pw. is degré(s)
cal.	calori	if bit 74 = 1 in pw. or if "pok"
c.-à-d.	sétadire	
cap., Cap.	capitale	if "pok"
cdt	comandan	
C.-du-N.	kôtdunor	
cf.	voir	
ch	chevaux	if bit 74 = 1 in pw.
chap.	chapitre	if bit 74 = 1 in fw.
ch.-l.	chef-lieu	
C^{ial}	commercial	
C^{ie}	compagnie	
cl	centilitre	if bit 74 = 1 in pw.
cm	centimètre	
col.	colonel	if "pok"
cont.	contenu	
cos.	cosinus	
coul.	coulomb	
cour.	couran	if "pok"
C^{te}	comte	
C^{tesse}	comtesse	
cv, CV	chevaux	if bit 74 = 1 in pw.
cv, CV	curiculommvité	in other cases
C.V.	curiculommvité	
dag	décagramme	if bit 74 = 1 in pw.
dal	décalitre	if bit 74 = 1 in pw.
dam	décamètre	if bit 74 = 1 in pw.
db.	décibel	

dép.	départ	if bit 74 = 1 in fw. or if fw. is à
dép.	départeman	in other cases
Desse	duchesse	
dg	décigramme	if bit 74 = 1 in pw.
div.	divèr	
dl	décilitre	if bit 74 = 1 in pw.
dm	décimètre	if bit 74 = 1 in pw.
DM	mark	if bit 74 = 1 in pw.
d°	ditto	only with superscript o
douz.	douzaine	
Dr	docteur	
DQ, D.Q.	derniékartié	
Ed	Edd	
éd.	édition	
édit.	édition	if "pok"
Env.	envoi	
env.	environ	
etc.	ètsétéra	
et/ou	éou	
EUA	étazunidamérique	
E.-U.	étazuni	
E.-et-L.	oeuréloir	
EV, E.V.	envil	
ex.	exemple	if "pok"
f	farad	if bit 74 = 1 in pw.
F	fran	if bit 74 = 1 in pw.
F	fahrenheit	if pw. is degré(s)
faub.	faubour	
fig.	figure	
g	gramme	if bit 74 = 1 in pw.
G	gauss	if bit 74 = 1 in pw.
Gal	général	
gal	général	
G-B, G.-B.	Grandbretagne	
h	heure	if pw. is une, deux, trois ... vingt-trois, vingt-quatre
h	habitan	in other cases
ha	ectare	if bit 74 = 1 in pw.
hm	ectomètre	
Hme	homme	
H.-Rh.	haurin	
HT	hors taxe	if pw. is franca, fran, or F
HT	hautetension	if pw. is volts
hW	hectowatt	
hz	hertz	
ibid.	ibidem	
id.	idem	if "pok"
in-4°	innqouarto	
in-8°	inoctavo	
J	joule	if bit 74 = 1 in pw.
J.B., J.-B.	janbatiste	
J-C, J.-C.	jézukri	
JF	jeunefille	

JH	jeunome	
J.J., J.-J.	janjak	
J.P., J.-P.	janpol	
jr	jour	
J3	jitroi	
k	kilo	if bit 74 = 1 in pw.
K	kèlvinn	if pw. is degré(s)
kg	kilo	
kHz	kiloertz	
kl	kilolitre	
km	kilomètre	
kmh, km/h	kilomètreur	
kV	kilovolt	
kW, kw	kilowatt	
kwh, kWh	kilowattheure	
l	litre	if bit 74 = 1 in pw.
larg.	largeur	
lat.	latitude	
lieut.	lieutnan	
lieut.-col.	lieutnancolonel	
loc. cit.	alendroicité	
log.	logaritme	if bit 74 = 1 in fw.
long.	longitude	if fw. is ouest, est, O, or E
long.	longueur	if "pok"
lu.	lumen	if bit 74 = 1 in pw. or if "pok"
m	mètre	if bit 74 = 1 in pw.
M.	monsieur	if "pok"
mA	milliampère	
M^al	maréchal	only with superscript al
Mat. Gr.	matièrgrasse	
max.	maximum	
M^e, M^es	maftre	only with superscript e/es
MF	miliondfran	if bit 74 = 1 in pw.
m.g.	matièrgrasse	
mg	milligrammme	
M^gr	monseigneur	
min.	minimum	
M^is	marquis	only with superscript is
M^ise	marquise	only with superscript ise
ml	millilitre	
M^lle	mademoiselle	
M^lles	mesdemoiselles	
MM.	messieurs	
mm	millimètre	
M^me	madame	
M^mes	mesdames	
M^r	monsieur	
mn	minute	
m/s	mètreparsegonde	
m/s^2	mètreçegonde	
ms	manuscri	
MTS	mètretonneçegonde	
mV	millivolt	

NB, N.B.	notabéné	
ND, N.-D.	notredame	
NDLR, N.D.L.R.	notdelarédaction	
N.-E.	norest	
NL, N.L.	nouvelllune	
N.-O.	norouest	
n°, n°s	numéro	only with superscript o/os
N.-S. J.-C.	notreçeigneurjézukri	
N-S J-C	notreçeigneurjézukri	
op.	opus	if "pok"
op. cit.	danlouvrajcité	
p., pp.	page	if bit 74 = 1 in fw.
P.	père	if pw. is le
p.a.	paran	
p. cent	poursan	
P.-de-C.	pâdcalais	
P.-de-D.	puidddôm	
p. ex.	parexemple	
pH	péash	
PL, P.L.	plènlune	
p.o.	parordre	
Pr, Prof.	professeur	if bit 75 = 1 in fw.
préc.	précédent	
pt	présidan	if bit 75 = 1 in fw.
q	quintaux	if bit 74 = 1 in pw.
Q.	question	if bit 74 = 1 in fw.
réf.	référence	
RSVP	répondéçilvouplè	
Rte	route	
R.V., R.-V.	renévictor	if bit 75 = 1 in fw.
R.V., R.-V.	rendévou	if bit 75 = 0 in fw.
s	seconde	if bit 74 = 1 in pw.
SAR	son naltèsroyal	if bit 75 = 1 in fw.
sc.	scène	
sdf	sandomicilfixe	
S.-E.	sudest	
sec., Sec.	section	if bit 74 = 1 in fw.
sec.	seconde	if bit 74 = 1 in pw. or if "pok"
S.-et-M.	sènémarne	
S.-et-O.	sènéoise	
sgt	sergent	
sin.	sinus	
SM, S.M.	samajesté	if bit 75 = 1 in fw.
S.-O.	sudouest	
SP	secteurpostal	if bit 74 = 1 in fw.
St	saint	
Ste	sainte	
Sté	société	
succ.	successeur	
suiv.	suivant	
SVP	silvouplè	
t	tonne	if bit 74 = 1 in pw.
tang.	tangente	

T.-et-G.	tarnégarone	
tél.	téléphone	
TSVP	tournéçilvouplè	
TTC, ttc, T.T.C.	toutetakskomprize	
Tt cft	touconfor	
V	volt	if bit 74 = 1 in pw.
var.	variante	
VDQS, V.D.Q.S.	vindélimité dekalitéçupérieur	
V.F., v.f.	versionfrançaise	
viz.	assavoir	if "pok"
V.O., v.o.	version originale	
vol.	volume	if "pok"
Vve	veuve	
vx	vieux	
W	watt	if bit 74 = 1 in pw.

APPENDIX V

TREATMENT OF HYPHENATED WORDS (except end-of-line hyphens)

Not all writers, printers, or dictionaries agree on which words should be written with a hyphen; for example, we find vert pomme and vert-pomme, mot outil and mot-outil, etc. Since problems occur with hyphenated words rather than with non-hyphenated words, our program has been written to treat correctly all hyphens, even those which are found only rarely.

We replace hyphens with a space and we set the features "prehyph" and "posthyph" to 1. The separated words of the hyphenated expression are taken through the normal transcription route (lexicon, roots, prefixes, transcription rules, etc.). Expressions such as the following are transcribed correctly after the hyphen has been replaced with a space:

hi-fi	casse-cul	sous-jacent	quatre-vingt-deux
tam-tam	chien-chien	vol-au-vent	quatre-vingt-huit
cow-boy	bêni-oui-oui	avant-bras	quatre-vingt-onze
abat-vent	miam-miam	électron-volt	quatre-vingt-neuf
cure-dent	outre-tombe	quatre-vingts	quatre-vingt-dix-sept
dix-sept	vif-argent	vice-présidents	un quatorze-juillet

On page 75, we gave the rule (Rule 21) which handles the cases of liaison in hyphenated expressions. The simplest case is when the first word ends with the letter t or the letter d. We have placed these exceptions on the idiom list: mot-outil, sourd-muet-aveugle, quatre-vingt-un, and the combinations with nord (e.g., nord-africain). The combinations with sud are also on the idiom list since the sequence /dt/ which is required for vendent-ils (/vãdtil/) is not acceptable for sud-est (/sydɛst/).

When the first word ends with the letter n, the liaison occurs only with a few words: en, bien, bon, non, plein, moyen; denasalization takes place with bon, non, plein, and moyen.

With words ending in s, x, or z, we had to write the rule so that the liaison would not take place in cases such as corps-à-corps, tentes-

abris, douces-amères, watts-heures. The list of idioms contains the exceptions which do not have a liaison in spite of the rule: saisies-arrêts, papiers-émeri, orangs-outan(g)s, guets-apens, tissus-éponges, trachées-artères and those which have a liaison in spite of the rule: vis-à-vis, quelques-un(e)s, tiers-arbitre/état/ordre, Basses-Alpes.

We have found only a very few hyphenated expressions in which a mute r precedes a hyphen and a vowel: laisser-aller, papier-émeri, and combinations with monsieur (e.g., monsieur-écologie); of those, only the first one is occasionally pronounced with an /r/ of liaison. For that reason, our rule does not consider words ending with the letter r.

Hyphenated expressions which present special problems have been added to the list of idioms: vingt-deux, chef-d'oeuvre, porc-épic, croc-en-jambe, arrive-t-il, va-t-en guerre, reviendra-z-à Pâques, etc.

APPENDIX VI: BORROWINGS, NEOLOGISMS, AND SLANG

A. Borrowings:

1. Many borrowings, old or recent, do not pose a problem because they do not require any special code in our program. Examples:

badge	ballast	baobab	budget	cacao	camping
cash	chott	ferry	fifty-fifty	forum	ghetto
grass	grog	harem	hasch	hi-fi	hobby
idem	julep	kilt	kolkhoze	kopeck	lifting
listing	macadam	marketing	match	maximum	milk bar
minimum	nabab	parking	party	passim	playback
pop corn	porridge	pressing	rallye	record	riesling
scotch	self	sexy	sketch	slip	snack-bar
snob	spaghetti	staff	standing	star	stop
studio	tandem	test	thalweg	tilt	totem
trolley	tunnel				

2. Among the borrowed words which cannot be transcribed correctly without some special code, there are some which are well established and which occur often enough in written French to justify that special code (e.g., use of the root table, special transcription rules). We cannot, of course, expect to give an acceptable transcription for each of the foreign words which may occur in today's newspapers[18]; the following list is more than adequate to meet the objectives of our program:

baby-sitter	bacon	baroud	basket	blitzkrieg
blockhaus	blue-jean(s)	bowling	boyscout	break
bulldozer	business	caïd	cake	charter
chewing-gum	chianti	clown	club	cocktail
coolie	corpus	couscous	crescendo	dead-heat
diesel	discount	drink	edelweiss	ersatz
fan	farniente	fez	flipper	flirt
foot	football	footballeur	footing	fortiori

gadget	gang	gangster	goal	golden
groom	hamburger	hydrofoil	igloo	interview
jazz	jean(s)	jet	job	jogging
kaiser	ketchup	khamsin	khmer	knock-out
knout	krach	kraft	kronprinz	kugelhof
leader	leitmotiv	looping	lunch	night-club
nurse	oued	pickpocket	pizza	placebo
poster	putsch	rumba	rushes	sandwich
scoop	shampo(o)ing	shoot	shooter	show
sit-in	sombrero	speakerine	speech	sprint
sprinter	squatter	steak	superman	suspense
tanker	tee-shirt	tramway	t-shirt	week-end
western	zoom			

Linguistic analysis is required to disambiguate the borrowed words jet, jean, poster, suspense from the French words with the same spelling.

Dictionaries do not always agree about the pronunciation of those borrowings. For example, the Petit Robert (1979) uses final /ɛr/ for tanker and final /œr/ for supertanker; the Petit Larousse (1981) uses final /œr/ for tanker and does not give a transcription for supertanker.

B. Neologisms:

The majority of the neologisms created during the last few years are transcribed correctly by our rules. Examples:

aéroglisseur	alcootest	aquaplaning	autopont
bureautique	calculette	dédensifier	deltaplane
désétatiser	déshachélémisation	déstabilisation	écolos
euromarcher	héliport	loubard	ludothèque
manif	maso	pédégé	phallocrate
pirater	pub	scanographie	sécu
sensass	sponsoriser	stagflation	tucard
visiophone	véliplanchiste	machmètre	

The following words are examples of neologisms which need to be included in our lexicon since our transcription rules do not transcribe them correctly:

abribus	altuglas	amniosynthèse	anéchoïque
bibliobus	bonus-malus	débitmètre	échographie
faitdiversier	métrobus	plexiglas	

The root for désatelliser/désatellisation has been included in our root table.

We have not included in our lexicon neologisms which have no chance of survival. For example, Le Canard Enchaîné (September 21, 1983) has an article entitled "Désimonisation"; it is only by reading the article that one finds out the word is to be pronounced /desi/ rather than /dezi/. The same satirical weekly, in its edition of March 21, 1984, uses globe-trotter as a verb to be pronounced with a final /e/ (Mais qu'est-ce qui fait

globe-trotter Tonton?); our lexicon contains only the noun globe-trotter pronounced with final /œr/.

C. Slang (Non-conventional French):

The vast majority of slang words are transcribed correctly by our transcription rules. In fact, a comparison of 1,000 random entries in the Petit Robert (1979) and in Cellard (1980) shows that the proportion of words not transcribed correctly by our rules is greater in the Petit Robert than in Cellard's slang dictionary (about 3% against about 2%). Most of the exceptions included in our lexicon belong to these groups:

1. final s is often pronounced: clebs, craps, doulos, etc.

2. vowel + m before consonant is not nasalized: fromgi, clamser, etc.

3. ch before consonant is often pronounced /ʃ/ instead of /k/: brichton, duchnoque, frichti, etc.

Occasional use of words in verlan (e.g., ripou for pourri, chébran for branché) or in trade jargons does not seriously affect our transcriber, but —of course— a whole text written in verlan (that is, with syllables à l'envers) or in the largonji des loucherbems (butchers' jargon) would not be transcribed properly since the table of endings, the root table, and the prefix table would give wrong information.

APPENDIX VII

PROCEDURES FOR THE ASSIGNMENT OF LINGUISTIC FEATURES

A. TYPES OF LINGUISTIC FEATURES NEEDED BY OUR PROGRAM:

1. One of the facts that make linguistic analysis difficult is that a given spelling may represent different words belonging to different grammar categories: for example, court can be a verb (il court), a noun (un court de tennis), an adjective (le plus court chemin), and an adverb (couper court à un entretien). Another reason is that the grammar category (or categories) of a prefixed word may not be the same as that/those of an unprefixed word: for example, porte is a noun and a verb, but déporte is only a verb.

But, for this program, we only need the level of accuracy that will allow us to make correct decisions about spellings which have several pronunciations (e.g., portions, président), length of breath-groups, liaisons, and intonation. We must, therefore, attempt to determine whether bien is a noun (no liaison) or an adverb (potential /n/ liaison), whether son is a noun (no liaison) or a determiner (potential /n/ liaison), whether en is a preposition or a pronoun, whether les is a determiner or a pronoun, etc., but there is no need to determine that bon, cher, clair, doux, dur, gras, gros, mauvais are adverbs in sentir bon, coûter cher, voir clair, filer doux, travailler dur, parler gras, parler gros, sentir mauvais.

Our analysis is adapted to our specific needs; for example, quelques, aux, mon are all placed in the category of determiners, miens, siens are placed in the category of nouns, and all the functions of qui are placed in the "subordinate" category.

2. A problem is posed by the many words which can function as noun and adjective, whether the meaning remains the same (e.g., russe, aveugle) or changes (e.g., bon, bas, bleu).

There are words which are usually perceived as nouns (e.g., ami, menteur, monstre, peuple) and which can also be used as adjectives: une maison amie, un enfant menteur, un dîner monstre, des manières peuple.

There are words which are usually perceived as adjectives (e.g., bas, bon, petit) and which are also used as nouns: une paire de bas, un bon du Trésor, la chienne et ses petits.

And there are words for which there can be hesitation. Are russe, présent, étranger, sauvage, sot, fou, ennemi perceived as adjectives which can also function as nouns or as nouns which can also function as adjectives?

In our program, from a pragmatic point of view, it often does not matter whether the adjective and noun features are correctly assigned. It never matters in the plural since we make the liaison whether the words are adjectives, nouns, or adjectives/nouns, but it can matter in the singular since we want a liaison in grand homme and we do not want a liaison in enfant italien. But, even so, the fact that there is no danger of wrong liaison when the first word ends with a letter other than d, g, n, r, s, t, x, or z or when the second word begins with a consonant sound reduces considerably the number of potential problem cases. Therefore, the table of endings and the lexicon can make simplifications which not only yield correct analyses in cases such as convalescent américain (no liaison), alsacien académicien (no liaison), malheureux ami (liaison), ancien adversaire (liaison), but also do not result in wrong liaisons even when the analysis is incorrect; for example, portique anglais is wrongly analyzed as adjective + noun, but it does not matter since no liaison is possible in this case; premier anniversaire is analyzed as adjective + adjective, but our program applies a rule (rule 20, page 75) which adds the liaison. The real problems are extremely rare and they are of a contextual nature; since we cannot perform the complex semantic analysis which, for example, would look at every usage of bon, grand, and premier to determine whether the meanings are good or bond, great/tall or nobleman, and first or prime minister, we have adopted the most frequent usage. Thus, we make the liaison in bon anglais (whether it means good English or English bond), in grand imbécile (whether it means tall imbecile or imbecile nobleman), in premier ambitieux (whether it means first ambitious man or ambitious prime minister). Since the combination savant aveugle is more often used as noun + adjective (with no liaison) than as adjective + noun (with liaison), we have placed savant in the lexicon as a noun to ensure that the analysis will treat aveugle as an adjective.

B. METHOD OF ASSIGNING LINGUISTIC FEATURES:

The linguistic features are assigned at three points in our program:

1. The features "cap" (bit 75) and "prnn" (bit 18) are assigned as described in Step 8 of Part V (page 18). These bits, once they are set, cannot be modified by any later step in the program.

2. Additional features are set while individual words go through the transcriber (pages 19-31).

3. Ambiguous features can be modified and additional features can be set by the linguistic analysis described on pages 52-65.

C. FEATURES ASSIGNED WHILE INDIVIDUAL WORDS ARE TRANSCRIBED:

1. When a word enters the transcriber, we set the features "sxzend" (words ending with s, x, or z), "dtend" (words ending with d or t), "nend" (words ending with n), "rend" (words ending with r), "gend" (words ending with g), "quote" (words which are preceded by guillemets), "prehyph" (words which precede a hyphen), "posthyph" (words which follow a hyphen).

2. Additional features are assigned when a word is found in the lexicon.

3. Additional features are assigned when the ending of a word is found in the table of endings.

4. A temporary plural feature ("temppl") is assigned if the word ends with the letter s at the time all final s letters are removed.

Assignment of features when a word is found in the lexicon.

a. Examples of words found during the first pass through the lexicon:

ennemi: this word is transcribed /ɛnmi/ and is marked "nn,sg".

boucher: since this word is found during the first pass, no prefix has been separated; the word is transcribed /buʃe/ and the ambiguous feature "nninf" is set IF the explicit feature "prnn" has not been previously set in Step 8 (page 18). The explicit features are those which appear in position 1 through 18 on our table (page 188) and our program does not allow an explicit feature to be changed once it has been set.

as: this ambiguous spelling is found in Group 5 of the lexicon; the feature "ambi" (bit 71) is set to 1 and the word will receive more analysis in a later part of the computer program (pages 59-65).

outres: this word is found during the first pass through the lexicon; it is transcribed /utrɛ/ and marked "nnvb"; outres must be in the lexicon since it does not belong to the same grammar category as outre. Our lexicon contains many such examples: biens, lits, mois, souris, tas, etc.

b. Examples of words found during the second pass through the lexicon:

contient: since this word is not in the lexicon, it cannot be found during the first pass; the word tient is found during the second pass after the prefix con has been separated and stored as /kɔ̃/; tient is transcribed /tjɛ̃/ and the features "v,sg" are set.

déboucher: since this word is not in the lexicon, it cannot be found during the first pass; the word boucher is found during the second pass after the prefix dé has been separated and stored as /de/; boucher is transcribed /buʃe/ and marked "inf" since the feature "npref" states that this is the category which must be used if a prefix has been separated.

désinfections: since this word is not in the lexicon, it cannot be found during the first pass; the word infections is found during the second pass after the prefix dés has been separated and stored as /dez/; since the spelling infections is ambiguous, feature "ambi" (bit 71) is set to 1 and the word is analyzed later as described on pages 59-65.

ennemie: since this word is not in the lexicon, it cannot be found during the first pass; the table of endings changes ennemie to ennemi and marks it "adjnvbpp,sg"; ennemi is then found during the second pass through the lexicon and the features become "nn,sg".

ennemis: since this word is not in the lexicon, it cannot be found during the first pass; when the program searches for final s, ennemis is changed to ennemi and marked "temppl" (temporary plural); ennemi is then found in the lexicon and the features "nn,pl" are set since the feature "temppl" is changed to "pl" when a word becomes classified as a word which can have a plural form (e.g., noun, adjective, past participle).

soient: since this word is not in the lexicon, it cannot be found during the first pass through the lexicon; it is marked "v,pl" when ient is replaced with i by the table of endings; soi is then found in the lexicon and is transcribed, but the features are not changed since explicit features assigned by the table of endings cannot be changed by the lexicon.

joaillier: since this word is not in the lexicon, it cannot be found during the first pass through the lexicon; it is marked "nn,m,sg" when the ending ier is replaced with ié by the table of endings; joaillié is then found in the lexicon and is transcribed; the features remain "nn,m,sg".

joailliers: since this word is not in the lexicon, it cannot be found during the first pass through the lexicon; it is marked "nn,m,pl" when the ending ers is replaced with é by the table of endings; joaillié is then found in the lexicon and is transcribed; the features remain "nn,m,pl" since a word which has already been marked plural cannot be changed to a singular assignment.

c. Examples of words found during the third pass through the lexicon:

ennemies: this word is not found during the first pass through the lexicon; it becomes ennemie after the final s is removed and the feature

"temppl" is set; ennemie is not found during the second pass through the lexicon; the table of endings changes ennemie to ennemi and the features become "adjnvbpp,temppl" (the feature "sg" is ignored if "temppl" is already set); ennemi is then found during the third pass through the lexicon and the features "nn,pl" are set since the feature "temppl" is changed to "pl" when the grammatical category becomes "nn".

balaies: this word is not found during the first pass through the lexicon; it becomes balaie after the final s is removed and the feature "temppl" is set; balaie is not found during the second pass through the lexicon; the table of endings changes balaie to balai and the features become "adjnvbpp,temppl" (the feature "sg" is ignored if "temppl" is already set); balai is then found during the third pass through the lexicon and the feature "nnvb" is set (the feature "temppl" is dropped since the category "nnvb" does not accept a plural feature).

Of course, a word may have to go through the lexicon more than three times; for example, antisuperhyperaéroenzymies (if it existed) would go through the lexicon seven times.

Assignment of features to words not found in the lexicon.

pore: since this word is not in the lexicon, since its ending is not in the table of endings, and since it does not have a final s, it completes the whole transcription process (/pɔr$/) without being assigned any linguistic feature.

pores, secours: such words receive the feature "temppl" when they lose their final s, but this feature is dropped when these words, not being caught by the table of endings or the lexicon, fail to receive a grammar category which allows singular and plural forms. The words end the transcription process with only the feature "exxend".

copie: this word receives the features "adjnvbpp,sg" when the ending ie is changed to i, but the feature "sg" is dropped when copi, not being caught by the lexicon, fails to receive a grammar category which allows singular and plural forms. The word ends the transcription process with only the feature "adjnvbpp".

copies: this word becomes copie after the final s is removed and the feature "temppl" is set; the table of endings changes copie to copi and sets the feature "adjnvbpp" (the feature "sg" is ignored if "temppl" is already set); the feature "temppl" is dropped when the following steps do not change "adjnvbpp" to a grammar category which allows singular and plural forms.

pleurs: this word becomes pleur after the final s is removed and the feature "temppl" is set; the table of endings assigns the feature "nn" to pleur and the feature "temppl" is changed to "pl" since pleur is now identified as a noun.

moellons: the root moel is found in the root table and moellons is

rewritten moillons. The table of endings changes moillons to moillon and assigns the feature "nnvb"; the following steps do not change this ambiguous grammar category to an explicit grammar category.

Châlons, Alger, Rabelais: we have not included in our lexicon the proper names which are transcribed correctly by our transcription rules. Therefore, proper names which, because of their position in the input string, have not been previously marked "prnn" (see page 18) may receive a wrong linguistic assignment from the table of endings: Châlons may be marked "nnvb", Alger may be marked "inf", Rabelais may be marked "v,sg"; the lexicon may also give wrong assignments (e.g., Clément may be marked "adj" by Group 1). The ambiguous features (e.g., "nnvb") may be corrected later by the linguistic analysis, but the explicit features (e.g., "inf") cannot be corrected; however, the few cases we have encountered in our tests have not posed a problem for the liaison process or the proper division into breath-groups. [Thus, there does not seem to be any justification for adding to our lexicon thousands of proper names which, although properly transcribed by our rules, may be assigned a wrong linguistic feature by our table of endings or our lexicon; this additional list would have to include all the proper names ending with one of 58 strings (e.g., ais, ai, ez, ix, ux, on, ers, er, ir, eur, ien, ard).]

D. FEATURES ASSIGNED AFTER THE STRING HAS BEEN TRANSCRIBED

After the words have been through the transcription program, some have complete linguistic features (e.g., serments "nn,m,pl,sxzend"), some have incomplete features (e.g., étais "nnvb,sxzend"), and some have no linguistic features at all (e.g., avocat, mordu).

No further analysis is necessary when the string is short enough not to require a division into shorter breath-groups and when there are no potential liaisons, even if the assignment of linguistic features is very rudimentary as in: Ce (dp) tendre (adjinf) domestique (nnvb) compte (nnvb) bien (nnadv).

However, input strings which do not require a division into shorter breath-groups and which do not have potential liaisons are rare. Most input strings must go through the linguistic analysis (pages 52-82).

APPENDIX VIII: LIST OF LINGUISTIC FEATURES

The table of linguistic features contains 112 entries. Entries 1 through 18 are for words which, in the input string being transcribed, have been identified as belonging to one specific category. Entries 19 through 39 are for words which have been identified as belonging to one of two categories; entries 40 through 47 are for words which have been identified as belonging to one of three categories; entries 48 through 51 are for words which have been identified as belonging to one of four categories.

We should note that the lexicon assigns the linguistic features to

individual words or idioms and that the ambiguous features are inclusive; marking mal as "nnadv" means that this word can function as a noun and as an adverb. The table of endings assigns the features to whole categories of words and the assignments are not necessarily inclusive; it marks the words ending with ons as "nnvb", meaning that the words can be verbs and/or nouns. In any case, our linguistic analysis (page 52) attempts to replace all ambiguous features (categories 19 through 51) with explicit features.

Our list of linguistic features is as follows:

1.	v	functioning as finite (personal) verb form: va, vas, vînmes
2.	imp	verb form functioning as imperative: vas-y
3.	inf	functioning as an infinitive: vêtir, croire
4.	nn	functioning as a noun: mouchoir, toit, le mien, chacun
5.	adj	functioning as a qualifying adjective: nouvel
6.	det	functioning as a determiner: une, quel, mon, cet
7.	prep	functioning as a preposition: sur, chez
8.	neg	functioning as a negation: ne, n'
9.	adv	functioning as an adverb: prudemment
10.	spr	functioning as a subject pronoun: je, il
11.	di	object pronouns which occur without a preposition: me, en
12.	disjpr	functioning as a disjunctive object pronoun: soi
13.	nb	functioning as a cardinal number: huit, dix
14.	sub	functioning as a subordinating word: dont, quoique
15.	ptp	functioning as a past participle: ouvert
16.	coord	functioning as a conjunction of coordination: et, ou
17.	prespar	functioning as a present participle: ayant, étant
18.	prnn	functioning as a proper noun: Faure, Paris
19.	adjnn	adjectives or nouns: malheureux
20.	nnvb	nouns or verb forms: pique
21.	nnadv	nouns or adverbs: bien
22.	detnn	determiners or nouns: ton
23.	pronpp	pronouns or past participles: tu
24.	dp	determiners or pronouns: le, leur
25.	ppr	prepositions or pronouns: en
26.	vbprep	verb forms or prepositions: entre
27.	prepant	prepositions or present participles: durant
28.	conjadv	conjunctions or adverbs: si
29.	adjadv	adjectives or adverbs: même
30.	allpron	pronouns which can be subject or object: lui, nous, elle
31.	vbpastpr	finite verb forms or past participles: peint, craint
32.	nnpastpr	nouns or past participles: vue, reçu
33.	nnant	nouns or present participles: habitant
34.	nnprep	nouns or prepositions: vers, avant, sous
35.	nninf	nouns or infinitives: goûter
36.	adjinf	adjectives or infinitives: tendre
37.	adjvb	adjectives or verb forms: prête, vive
38.	prepadj	prepositions or adjectives: sauf
39.	adjpp	adjectives or past participles: pressé, gelé
40.	nprepant	nouns, prepositions, or present participles: devant
41.	adjnnvb	adjectives, nouns, or verb forms: vide
42.	adjnnpp	adjectives, nouns, or past participles: mort
43.	adjnnant	adjectives, nouns, or present participles: tournant

44.	nnadjadv	nouns, adjectives, or adverbs: fort
45.	nnvbadv	nouns, verbs, or adverbs: point
46.	nnvbprep	nouns, verbs, or prepositions: outre
47.	vbadvpp	verbs, adverbs, or past participles: plus
48.	ppadnant	prepositions, adjectives, nouns, or present participles: pendant
49.	adjnadvb	adjectives, nouns, adverbs, or verbs: court, faux
50.	adjnvbpp	adjectives, nouns, verbs, or past participles: distrait
51.	adndp	adjectives, nouns, determiners, or pronouns: nul
52.	m	masculine
53.	f	feminine
54.	ag	both genders: livre, voile
55.	sg	singular
56.	pl	plural
57.	an	both singular and plural forms: tas, noix, nez
58.	temppl	temporary plural feature (word ending with s or x)
59.	sxzend	word ending with s or x or z: gros, faux, nez
60.	dtend	word ending with d or t: grand, petit
61.	nend	word ending with n: ancien
62.	rend	word ending with r: premier
63.	gend	word ending with g: long
64.	npref	use linguistic features preceding "npref" if a prefix has been separated; if not, use features following "npref"
65.	z	potential liaison with /z/: vous, deux, chez
66.	n	potential liaison with /n/: on
67.	t	potential liaison with /t/: dont, sont, quand
68.	p	potential liaison with /p/: trop
69.	r	potential liaison with /r/: since our present program does not make the liaison with er infinitives, we do not use this bit; this slot is provided for future versions of our program (e.g., poetry) in which the er infinitives would make the liaison.
70.	tn	must go through special process: six, cinq
71.	ambi	spellings which have two or more pronunciations: fils
72.	prehyph	words which precede a hyphen: prenez-en
73.	posthyph	words which follow a hyphen: prenez-en
74.	rwnb	number changed to graphemic form
75.	cap	words which begin with a capital letter
76.	quote	words preceded by guillemets

Categories 77 through 112 are for lists of words which are necessary to refine the grammatical analysis described in Chapter VII and to solve ambiguities (e.g., fils). These lists are in Appendix XVII, pages 204-207.

Thus, 112 bits of information are available for each word of the input string. These bits can be set as the word goes through Step 8 (page 18), as the word enters the transcriber (page 19), and as it goes through the lexicon, the table of endings, the search for final s, and the linguistic analysis. For example, leur completes the phonemic transcription with these bits set to 1: 24, 54, 62, 79, 87, 88, 92, 94, 95, 100, 105, 108; if the linguistic analysis shows that leur is being used as a determiner, bit 24 is reset to 0 and bit 6 is set to 1; if the analysis shows that leur is a pronoun, bit 24 is reset to 0 and bit 11 is set to 1.

APPENDIX IX: ROOT TABLE

The 228 roots used by our program are listed below. In each of the three groups, the first column lists the roots which are to be rewritten; the second column gives the rewritten roots; the third column lists conditions under which the rewriting is done: "2" if the rewriting is done only when the following character is a consonant; "4" if the rewriting is done only when the following character is a vowel.

When a root ends with a vowel, the modification is also made if the vowel is accented; for example, the entries chample, reste also serve for champlè, restê, restè.

In the rewritten roots, the string wzx is transcribed as /ã/ by our transcription rules and the string wxz is transcribed as /gn/.

accourr	acourr		bisili	biçili		désulf	déçulf				
acquerr	aquerr		bisul	biçul		désouf	déçouf				
acupun	akupon		boycott	bwwkott		désyn	déçyn				
agna	awxza		brach	brak		détaill	dettaill				
agno	awxzo		broncho	bronko		diagno	diawxzo				
aiguill	égüill		cauc	koc		dicho	diko				
alcoo	alko		cautér	kotér		distill	diçtil				
als	alz	4	cautèr	kotèr		dys	diç				
amygd	amid		chample	chanle		enamour	wzxnamour				
anach	anak		chiro	kiro		enchn	enschn				
appendic	apindic		chlorosu	kloroçu		encht	enscht				
archa	arka		chol	kol		enhard	wzxard				
archéo	arkéo		chor	kor		enharm	wzxnarm				
argu	arguu	4	clams	clammç	4	enharn	wzxarn				
asep	açèp		cognat	kowxzat		enher	wzxnèr				
asex	açèkç		cognit	kowxzit		enivr	wzxnivr				
asoci	açoçi		conch	konk		enorgu	wzxnorgu				
assen	assén	4	contres	kontreç	4	équerr	ékerr				
asthm	açm		cos	koç	4	extrai	èkçtrai				
asy	açy		couenn	koin		extray	èkçtray				
augm	ogm		crams	crammç	4	érenti	érançi				
augu	ogu		damn	dân		eû	u				
aumð	omð		décemvi	déçèmvi		eure	u're	2			
ausc	oçk		désacra	déçakra		euss	uç				
aust	ost		désanct	déçankt		exempt	ègzant				
auth	oth		désatell	déçatell		exs	èkç				
auti	oti		déségr	déçégr		fayo	faillo				
aux	okç	4	désensi	déçançi		flirt	fleurt				
bacill	baçil		désexual	déçexual		gaî	gué				
balbuti	balbuçi		désign	désign		gaie	gué				
bals	balz	4	désili	déçili		gnat	wxzat				
bay	baill	4	désolid	déçolid		gneiss	wxzèç				
benz	binz		désorpti	déçorpçi		gnom	wxzom				
bien	biin		desquam	dèçkouam		gnos	wxzos				
bisegm	biçègm		désud	déçud		guadelou	gouadelou				
biséri	biçéri		désue	déçue		homosex	omoçèkç				
bisex	biçèkç		désuint	déçuint		ichor	ikor				

igni	iwxzi	quadra	kouadra	ressuy	réçuy
impatien	impaçien	quadrill	kadrill	restab	restab
initi	iniçi	quartz	kouartç	restat	restat
insatiab	inçaçiab	quatern	kouatèrn	resta	rèsta
instill	inçtil	raugm	rogm	reste	rèste
interro	intairo	recev	recev	restit	rèstit
interru	intairu	reç	reç	restrei	rèstrei
isch	içk	referen	référin	restrict	rèstrict
isthm	içm	récogni	rékowxzi	retour	retour
jers	jèrz 4	rect	rèkt	revolv	révolv
lingua	lingoua	regard	regard	rhynch	rink
lingui	lingüi	regret	regret	rouenn	roin
machia	makia	réglem	règlem	saoul	çoul
manich	manik	remb	rwzxb	secon	çegon
mauvai	movai	remm	rwzxm	sempiter	çinpitèr
métamor	mettamor	remp	rwzxp	shoot	chout
mill	mil	renamour	rwzxnamour	solenn	çolan
moel	moil	renher	rwzxnèr	souhait	çouèt
oignon	ognon	renivr	rwzxnivr	sprint	çprinnt
orch	ork	renon	renon	stagn	çtawxz
oscill	oçil	renorgu	rwzxnorgu	subsist	çubziçt
parasit	parazit	ren	rwzx 2	substanti	çubçtançi 4
patien	paçien	repos	repos	susur	çuçur
pétiol	péçiol	rept	rèpt	tachéo	takéo
pignora	piwxzora	rescind	réscind	tachy	taki
poêl	poual	résect	réçèkt	titill	titil
présal	préçal	réséqu	réçék	trach	trak
présan	préçan	résèqu	réçèk	tranquill	trankil
présé	préçé	respect	rèçpect	transce	trançe
présign	préçign	respir	rèçpir	trichi	triki
présuc	préçuc	resplen	rèçplen	tricho	triko
présupp	préçup	respons	rèçpons	trise	triçe
présy	préçy	resquill	rèçkill	trisul	triçul
progn	prowxz	ressaie	réçaie	trisy	triçy
prompt	pront	ressass	reçâç	view	viouv
prosi	proçi	ressay	réçay	vraisem	vrèçan
prosy	proçy	ressembl	ressembl	zon	zôn
psycha	pçika	ressemel	ressemel		
psychia	pçikia	ressui	réçui		
pyrosulf	piroçulf	ressusc	réçuç		

APPENDIX X: PREFIX TABLE

The 65 prefixes used by our program are listed below. The numbers indicate the conditions under which the prefixes are separated:

1 if followed by a letter other than s (prénom, prééminent). The problems posed by pré + s are solved with the root table and the lexicon.

2 if followed by a consonant other than n (contenu, emmener). This is to avoid the problems which would occur with words such as connivence, entrées, métairie, miniature, paraître, parangon, union.

3 if followed by s + vowel. This is to help transcribe words such as dessers, dessaouler.

4 if followed by a vowel or h (désir, déshydrater). Cases such as désolidariser are handled by the root table.

5 if followed by a consonant other than s (détenir).

6 if followed by é. This is to handle the problem posed by the forms of the verb différentier and its derivatives. In our PLATO program, the root différenti cannot be placed on the root list because it is too long; we place diff on the prefix list and érenti on the root list.

7 separated as /ɛ̃/ if followed by a consonant other than n or h (incertain); as /in/ if followed by a vowel (inexorable); not separated otherwise (innocent, inhumain).

8 /trɑ̃z/ if followed by a vowel or h (transistor, transhumance); /trɑ̃s/ otherwise (transsibérien).

9 not separated if followed by a vowel. This is to simplify the code required to transcribe words such as réalité, Emotion (see footnote 7).

aéro	aerɔ		hypno	ipnɔ		photo	fɔtɔ	
anti	ɑ̃ti		hypo	ipɔ		physio	fizjɔ	
aqua	akwa		idio	idjɔ		poly	pɔli	
archi	arʃi		inter	ɛ̃tɛr		post	pɔst	
auto	ɔtɔ		in	ɛ̃	7	pré	pre	1
bio	bjɔ		macro	makrɔ		proso	prɔzɔ	
circum	sirkɔm		méta	meta	2	pseudo	psødɔ	
con	kɔ̃	2	micro	mikrɔ		psycho	psikɔ	
crypto	kriptɔ		mini	mini	2	quadri	kwadri	
des	de	3	mono	mɔnɔ		quadru	kwadry	
dés	dez	4	morpho	mɔrfɔ		radio	radjɔ	
dé	de	5	mos	mɔz	4	re	rə	9
diff	dif	6	moz	mɔz	4	simili	simili	1
électro	elɛktrɔ		multi	mylti		sub	syb	1
electro	elɛktrɔ		néo	neɔ		super	sypɛr	
em	ɑ̃	9	omni	ɔmni		supra	sypra	
ennéa	enea		ortho	ɔrtɔ		télé	tele	
entre	ɑ̃trə	2	ostéo	ɔsteɔ		thermo	tɛrmɔ	
extra	ɛkstra		oxy	ɔksi		trans	trɑ̃z	8
hendéca	ɛ̃deka		para	para	2	ultra	yltra	
hydro	idrɔ		penta	pɛ̃ta		uni	yni	2
hyper	ipɛr		philoso	filɔzɔ				

The strings mos, moz, philoso, proso have been placed on the list of prefixes in order to keep open /ɔ/ before /z/ in words such as mosaïque, mozarabe, philosophe, prosodie.

We must place in the lexicon the prefixes which, in isolation, would not be transcribed correctly by our rules: aqua, ennéa, hyper, inter, penta, photo, psycho, quadri, quadru, trans.

APPENDIX XI

TABLE OF ENDINGS: MODIFICATIONS AND ASSIGNMENT OF FEATURES

gents	jan	adj,m,pl	rut	ru	v,sg
cents	çan	adj,m,pl	ionne	ionne	v,sg
ents	an	nn,m,pl	eur	eur	nn,sg
ément	éman	adv	ique	ike	adj,ag,sg
emment	eaman	adv	ic	ik	nn,m,sg
amment	aman	adv	ien	ien	nn,m,sg
oiement	oiman	nn,m,sg	vienne	viène	v,sg
iquement	ikman	adv	tienne	tiène	v,sg
alement	alman	adv	ienne	iène	nn,f,sg
ellement	elman	adv	ante	ante	adj,f,sg
eusement	euzman	adv	ance	ance	nn,f,sg
gement	jman	nn,m,sg	sme	sme	nn,m,sg
onnement	onnman	nn,m,sg	um	um	nn,m,sg
ssement	sman	nn,m,sg	age	age	nn,sg
chement	cheman	nn,m,sg	au	au	nn,m,sg
ivement	ivman	adv	aire	aire	adj,ag,sg
ement	eman	nnadv	ère	ère	nn,sg
scent	ssan	nn,m,sg	euse	euse	nn,f,sg
guent	gue	v,pl	ble	ble	adj,ag,sg
uent	u	v,pl	ard	ar	nn,m,sg
ient	i	v,pl	al	al	adj,m,sg
ent	e	v,pl	ot	o	nn,m,sg
ais	ai	v,sg	ant	an	adjnnant
ai	é	v,sg	in	in	nn,sg
ez	é	v,pl	iste	iste	nn,sg
ix	i	nn,ag,an	ième	ième	adj,ag,sg
ux	u	adjnn	oire	oire	nn,sg
ations	assion	nn,f,pl	ou	ou	nn,m,sg
ctions	cssion	nn,f,pl	an	an	nn,m,sg
ptions	pssion	nn,f,pl	if	if	adj,m,sg
tions	tion	v,pl	erie	eri	nn,f,sg
stion	stion	nn,sg	ogie	ogi	nn,f,sg
xtion	xtion	nn,f,sg	ès	èxx	nn,m,an
tion	ssion	nn,f,sg	ss	xx	nn,an
ons	on	nnvb	gue	gue	adjnnvb,sg
on	on	nn,sg	ée	é	adjnvbpp,sg
ers	é	nn,m,pl	cratie	craçi	nn,f,sg
ier	ié	nn,m,sg	ie	i	adjnvbpp,sg
er	é	inf	que	ke	adjnnvb,sg
oir	oir	nn,m,sg	ue	u	adjnvbpp,sg
air	air	nn,m,sg	nc	n	adjnnvb,sg
irs	ir	nn,m,pl	ing	ign	nn,m,sg
ir	ir	inf	ng	n	nn,m,sg
ndre	ndre	inf	ale	ale	adjnnvb,sg
ont	on	v,pl	elle	elle	adjnnvb,sg
end	ean	v,sg	ile	ile	adjnnvb,sg
ra	ra	v,sg	ole	ole	adjnnvb,sg
ait	ai	v,sg			

APPENDIX XII: ALGORITHM FOR INITIAL /ˇ/

A. In French, there are several hundred words which, although beginning with a vowel sound or semivowel sound, do not behave like other words beginning with a vowel or semivowel sound.

The majority of those words begin with the letter h (improperly called aspirate h), but about fifty begin with a vowel letter and about twenty begin with the letter w. We will use the term hiatus word to designate such words. Examples:

haut	honte	ah	uhlan	oui	ululer
haricot	hurler	onze	yaourt	watt	

Hiatus words behave as follows:

1. They do not allow the liaison process:

en haut	is pronounced	/ɑ̃o/
nous hurlons	is pronounced	/nuyrlɔ̃/
un oui	is pronounced	/œ̃wi/
deux watts	is pronounced	/døwat/

(The /z/ and /t/ in dix-huit and vingt-huit are not liaisons since /z/ is pronounced in dix-neuf and /t/ is pronounced in vingt-neuf.)

2. They do not normally allow the fall of /ə/:

le onze avril	is pronounced	/ləɔ̃zavril/
ce yaourt	is pronounced	/səjaurt/
quelle honte	is pronounced	/kɛləɔ̃t/

However, in some phrases, some people replace /ə/ with a very short pause (e.g., /kɛl.ɔ̃t/) or drop the /ə/ (e.g., c'est le huit, à pareille hauteur pronounced /selɥit/, /aparɛjotœr/). Our program keeps /ə/ except when huit is preceded by a hyphen (e.g., trente-huit).

3. They do not usually allow linking:

il hait	is pronounced	/il.ɛ/

as opposed to:

il est	pronounced	/ilɛ/

However, linking generally takes place in par hasard, leur huit enfants, vingt-huit, super-huit, un pareil héros, etc. Our program allows linking when huit is preceded by a hyphen, when huit is preceded by a determiner, and in par hasard (placed on the idiom list of the lexicon).

B. Un is a hiatus word when used as a number rather than an article:

chapitre un /ʃapitrəœ̃/ cent un /sɑ̃œ̃/

The cases in which /ə/ is kept before un have been included in the idiom list of the lexicon.

C. A hiatus word remains a hiatus word when used with a prefix. Compare:

en + herber	is pronounced	/ɑ̃nɛrbe/
en + hardir	is pronounced	/ɑ̃ardir/

D. It is essential that all hiatus words be correctly marked. Those which begin with a letter other than h have been included in the lexicon. For those which begin with the letter h, we assume that they are hiatus words unless the words themselves or the roots appear on one of these lists:

1. Lexicon list: These words have been placed in the lexicon because they present special problems (e.g., hibiscus, haïtien) or because the root is used in only one word (e.g., hafnium, haliple).

h	habeas	habitus
hacienda	hafnium	haggis
Haïti	haïtien	haïtienne
haleine	halicte	halieutique
haliotide	haliple	hallali
haltère	hamadryade	hamadryas
hamamélis	hapax	harpagon
haruspice	hast	Hawaii
hawaiien	hawaiienne	Hawaï
hawaïen	hawaïenne	heiduque
Hendaye	hendayen	hendayenne
hendyadis	hendyadyin	henry
heptasyllabe	heptasyllabique	herpès
hetman	hexasyllabe	hexasyllabique
hécatombe	hélépole	hélodée
hérédosyphilis	hérédosyphilitique	hétérosexualité
hétérosexuel	hétérosexuelle	hévéa
hiatus	hibiscus	hidalgo
hidrosadénite	hier	hièble
himation	hindi	hinterland
hippodrome	holmium	hombre
homothétie	homuncule	honneur
honoris	hoplite	horaire
hui	humeur	humérus
humour	humus	hurluberlu
hymen	hypnoses	hypospadias

The word haste is in Group 5 of the lexicon (ambiguous spellings).

2. List for transcription rules: The following strings are transcribed by our transcription rules without initial /ˇ/:

habi	hagio	halen	halèn	halitu	halluci
halog	haloph	haltè	hameç	haplo	harmo
hebdo	héb	hect	héd	hég	hel + C (1)
héli	hém	hend	hépa	hepta	héral

herb	hercu	hercy	hérẽ	hẽrit	herm
hẽroĩ	herpẽ	hertz	hẽt	heur (2)	hex + V
hẽz	hiata	hiber	hiẽm	hiẽrat	hiẽro
hilar	hilo	himalay	hin	hipp (3)	hir
hisp	hist	hitl	hiv	hoir	holo
homẽ	homi	homm	homo	honnẽ	hono
hõpi	hordẽ	horiz	horl	hormo	horo
horr	hort	hos	hõt	hoz	huil
huis (4)	huĩt	humai	human (5)	humb	humect
humẽr	humi	humor	hy		

(1) except hello (placed in Group 3 of the lexicon)
(2) except the string heurt (placed, in the transcription rules, before the string heur)
(3) except hippie, hippy (placed in Group 3 of the lexicon)
(4) except in huis clos (placed on idiom list of the lexicon)
(5) except humant (placed in Group 3 of the lexicon)

3. The words beginning with hendéca, hydro, hyper, hypno, and hypo are handled by the prefix table.

4. The words beginning with homosex are handled by the root table.

The preceding lists have been established with the help of seven dictionaries; there is general agreement except for a very small number of words (e.g., hélas, hiéroglyphe, hindi, hyène, hayon, huitante). Haïti and haïtien present a strange case: the Petit Larousse (1981) considers that haïtien is a hiatus word and that Haïti is not while the Petit Robert (1979) does exactly the reverse. In all those cases, we have adopted the pronunciation favored by the majority.

Since intervocalic s has already been changed to z when we reach this part of the computer program, the strings hẽz and hoz represent the roots hẽs (e.g., hésiter) and hos (e.g., hosanna).

English borrowings (e.g., hall, home) are in Group 3 of the lexicon.

APPENDIX XIII: DISTRIBUTION of /a/ and /ɑ/

The a/ɑ contrast, for those speakers who make it, is phonetic (fable with [ɑ] vs. table with [a]), phonological (pâte with /ɑ/ vs. patte with /a/), and semantic (in the 1979 edition of the Petit Robert, all the words in which the root bas means low or base have /ɑ/: basse, basset, basson, basique, basophile, etc. while all the others have /a/: bassin, basin, basilic, basalte, etc.).

P. Fouché, in his Traité de prononciation française (1956), describes the pronunciation generally used by educated native Parisians in the first half of this century. His description is very detailed (pages 56-63 for stressed syllables, pages 84-90 for unstressed syllables) and extremely complex (for example, the pronunciation of the noun taille depends on its meaning).

Fouché is clearly in favor of maintaining the distinction a/ɑ, but he admits (page 56) that "la répartition de [a] et [ɑ] dans le français correct d'aujourd'hui est assez délicate à établir." Furthermore, the following statements show that he is aware his description holds true for only a small proportion of French "locuteurs" and that it will remain true for only a short time:

> page 62: On peut encore prononcer un [ɑ] bref dans les mots bois (substantif), mois, pois..... (+ a list of about 50 words).
>
> page 62: Partout ailleurs on ne prononce plus guère que [a].
>
> page 85: L'[ɑ] ne fait que se survivre dans baron, brasiller, carrosse.... (+ a list of about 40 words).
>
> page 85: L'[ɑ] a pour ainsi dire disparu dans abrasif, avocaillon, basalte.... (+ a list of about 50 words).

Pierre Delattre (1949) considered that the contrast a/ɑ was doomed:

> page 43: The distinction between front a [a] and back a [ɑ] is fast disappearing in the present generation. A somewhat central a is coming to be used for all purposes. This central a may vary considerably from front to back, giving it a wider range of color variations than any other vowel.

but Delattre (1957) seemed to recommend that students should learn to make the distinction:

> page 148: C'est au 18e siècle, avec l'antériorisation des formes verbales en âmes et âtes, que les A postérieurs ont gravement perdu du terrain. Depuis, ils continuent à décroître en nombre. Dans la période contemporaine de Passy à Fouché, c'est-à-dire en quelque cinquante ans, ils semblent avoir réduit de moitié. Parallèlement, on constate un rapprochement phonétique des deux A dans le français cultivé. Devant cette double attaque, statistique et phonétique, l'A postérieur continue-t-il à battre en retraite? Les jugements sur le présent sont impossibles; à plus forte raison sur l'avenir. Il ne serait pas prudent de conseiller aux étudiants que le nombre des [ɑ] rebute encore de remettre leurs études de cinquante ans.

The Petit Robert (1979) states on page xxii that "on observe une régression de la distinction entre [a] et [ɑ]: pâte [pɑt] tend à se rapprocher de patte [pat]". It is therefore surprising that this dictionary should transcribe all words ending in ation as [ɑsjɔ̃], a pronunciation which is only rarely heard today; in 1956, Fouché already stated (page 85) that "dans nation, ration, station, l'[ɑ] est en train de céder la place à [a]."

The most categorical statement is made by A. Lerond (1980), page xii:

> "Mais, à Paris même, chez les personnes nées en 1950 ou postérieurement, la distinction entre [a] et [ɑ] ne se fait plus, la voyelle étant d'ordinaire réalisée avec un timbre intermédiaire, plus proche

> néanmoins de [a] que de [ɑ]. Cette réalisation paraît être actuellement la plus usuelle dans la majeure partie de la francophonie .... mais, puisque certains Parisiens continuent à opposer [ɑ] tonique à une voyelle moins postérieure, nous ajoutons, après la prononciation neutre, celle où un [ɑ] est apparu dans nos enquêtes -- précédée de la mention "vieilli Paris". De la sorte, on assigne au [ɑ] "traditionnel" la place qu'il occupe dans la réalité de la langue, celle d'un archaïsme d'extension réduite."

Our research in France, during the summer of 1983, does not substantiate Lerond's statement. While it is true that the a/ɑ situation is not what it used to be, there are French natives born after 1950 who have a clear a/ɑ contrast and we believe that our program should attempt to reproduce what the "average" French native hears on radio, on television, and in movie theaters as well as in his daily conversations.

Since the French natives who have some form of a/ɑ contrast do not agree on 1) the timbre of the vowels and the extent of the contrast and 2) the list of the items which use /ɑ/, it is impossible to write transcription rules which would please everyone; the following is a compromise which, we hope, will prove acceptable to most people:

1. Our transcription rules yield /ɑ/ for the following strings:

   a) strings damn, passe, tasse in all positions

   b) strings basse, bataill (second a), class, fabl, gagn, raill, sabl, taill in initial position or preceded by one of our prefixes

   c) strings abre, adre, afle, ase, avre, aze, racle in final position (including forms with final s and nt)

   d) all â except: final ât, âmes, âtes, but not including: appât, bât, dégât, mât, âmes, blâmes, infâmes, pâmes, appâtes, bâtes (and prefixes), gâtes, hâtes, mâtes (and prefixes), pâtes (and prefixes), tâtes (and prefixes)
   and except: bâbord, bâbordais, bâbordaise

2. With those transcription rules, the number of words which remain to be placed in the lexicon is relatively small; the list contains some words which would be acceptable with /a/, but —since they have to be in the lexicon for other reasons— using /ɑ/ does not require any extra computer memory. The list is as follows:

ah	Algésiras	antibois	appas	as[1]	Assas
atlas	Augias	bah	bas	bâta[2]	Blas
bois[3]	cabas	cadenas	cas	casse	Chasles
Chastellain	chocolat	coutelas	crabe	cycas	dégras
Delmas	diable	Dukas	Duras	encas	endiable
ensable	entrelacs	fa	frimas	gable	galetas
gars	gaz	glas	gras	grasse	grassement
hamadryas	haras	hautbois	Havas	havresac	hélas[4]

jazz	Jonas	Juliénas	k	kvas	kwas
lacs[5]	Ladislas	las	lasse	lias[6]	lilas
Madras	mas	Mathias	Ménélas	merdas	mois
noix	nostras	paille	pas	Pasdeloup	Pelléas
pois	poix	Praslin	psoas	rail	rasibus
ras[7]	raz	repas	Stanilas	tas	tétras
trépas	trias[8]	trichomonas	trois	Troisgros	upas
vasistas	Venceslas	vindas	wombat		

(1) if noun, (2) if synonymous with "stupide", (3) if noun, (4) if adverb, (5) if synonymous with "piège", (6) if noun, (7) if Ethiopian title, (8) if noun

3. On speech synthesizers, experiments should determine which formant settings provide the a/ɑ contrast which most people find acceptable for male voices and for female voices.

APPENDIX XIV: The consonant /ɲ/

Alain Lerond (1980) states on page xiv:

> "La consonne [ɲ] est une ancienne consonne, articulée en portant la langue antérieure sur le milieu du palais dur, qui permettait, par exemple, de distinguer la séquence au pignon [opiɲɔ̃] du mot opinion [opinjɔ̃]. Depuis plusieurs générations déjà, on tend à lui substituer le groupe consonantique [nj]. Subsistant dans le français neutre comme dans les français régionaux, [ɲ] devait figurer dans notre dictionnaire, à condition d'être signalé comme un archaïsme."

We believe that Lerond goes too far in his assessment of /ɲ/. Our own study does indicate that only a minority of young French speakers possess the consonant /ɲ/, but —if we consider all age groups— /ɲ/ is still used by a majority of the French population. The term "archaic" is not justified (at least, not yet) and the most recent Robert and Larousse dictionaries still use /ɲ/ in their transcriptions.

APPENDIX XV: TRANSCRIPTION OF oe and œ

A. Since some typesetting equipment and most typewriters do not have the grapheme œ and use oe instead[19], our program must be able to handle inputs such as oeil, soeur, oestre, foetus, coeliaque, désoeuvrement as well as œil, sœur, œstre, fœtus, cœliaque, désœuvrement; it must also be able to distinguish that alternate form of œ from the "legitimate" o + non-accented e (e.g., coefficient, moelle, hypoesthésie) and from o + accented e (e.g., poète, poêle, proéminent, coéducation, Noël, canoë).

Since our code must be able to handle all these occurrences of oe, the code for œ would be redundant; it is simpler (see page 13) to rewrite all occurrences of œ as oe at the very beginning of the program.

To our knowledge, no one has compiled and published a complete list of

the French words which contain the grapheme œ. The list we have drawn up is probably not complete, but it is large enough to allow us to establish accurate transcription rules.

A complete list of the words containing a legitimate oe (not an alternate spelling of œ) does not exist either. The list we have established indicates that, except for the strings moell and groenland, rare words like moere, boette, and proper names like Ploemeur, the "legitimate" oe occurs only when a prefix ending with o precedes a root beginning with e (e.g., coexistence, gastroentérite). Some printers place a hyphen between the prefix and the root (e.g., gastro-entérite); this, of course, removes the problem of determining whether oe is an alternate spelling of œ.

Using a list of 197 prefixes ending with o, we found only the following combinations in the Petit Robert (1979) and the Petit Larousse (1981):

- prefix co + roots found in coefficient, coercible, coexister, coenzyme, coextensif
- prefix électro + root of encéphalogramme
- prefix gastro + root of entérite
- prefix hypo + root of esthésie
- prefix leuco + root of encéphalite
- prefix neuro + root of endocrine

but, clearly, words such as coeffet, neuroenchaînement, coexaminateur, monoexistence can be used and our program should transcribe them correctly.

Our computer program treats the oe spelling as follows:

1. Some words have to be placed in the lexicon. They are:

- Words which have pronunciation problems in addition to the problem posed by oe. Examples (we underline the letters causing the problems):

boeufs	choeur	moeurs	oersted	oeufs	stoechiométrie
Chloé	foetus	oenothera	oestrus	phoenix	stoechiométrique

- Words for which it is not practical to make up rules (there are not enough words to justify a rule). Examples:

Boers	canoë	Goethe	loess	Ploemeur
boëtte	foëne	Groenland	oerstite	roentgen

- Monoecie and monoecies since replacing œ with oe creates words for which we must not allow prefix separation: mono + ecie(s).

2. Two roots have been placed in the root table: moell, poël.

3. For the other words, we apply these rules (see pages 37, 39, 44):

a. In the string o + accented e, o is transcribed as /ɔ/: poète, coédition, Noël, proéminence, etc.

b. In the string oeil, oe is always transcribed with /œ/: oeil, oeillade, oeillère, oeillet, oeilleton, oeilletonner, etc.

c. The string oeu is transcribed with /ø/ in final oral position of words (voeu, voeux, noeud, noeuds, etc.) and transcribed with /œ/ in all other cases (soeur, boeuf, oeuvre, coeur, manoeuvrable, désoeuvré, etc.).

d. In all remaining cases, o + unaccented e is followed by a consonant letter. Since the number of words with a "legitimate" oe is greater than the number of words in which oe is an alternate spelling of œ and since the number of words with œ is finite while it is possible to create more "legitimate" words by combining an -o prefix with a root beginning with unaccented e, it is preferable to write the rules as follows:

- Transcribe coe as /se/ in coel + vowel: coeliaque, coelentéré, etc.
- Transcribe coe as /se/ in coen + vowel: coenure, coenesthésie, etc. (A problem would occur if the prefix co was used without a hyphen with the roots of enivrer or enhardir.)
- Transcribe foe as /fe/ in foet + vowel: foetal, assa-foetida, etc.
- Transcribe poe as /pe/ in poec + vowel: poecile, poecilotherme, etc.
- Transcribe oe as /e/ in oecum + vowel: oecuménique, oecuménisme, etc.
- Transcribe oe as /e/ in oed + vowel: oedipien, oedème, oedicnème, etc.
- Transcribe oe as /e/ in oen + vowel: oenologue, oenanthe, oenothère, etc.
- Transcribe oe as /e/ in oesoph: oesophage, oesophagien, etc.
- Transcribe oe as /ɛ/ in oestr + vowel: oestrogène, oestre, etc.
- In all other cases, we have a "legitimate" oe and the letter o is transcribed as /ɔ/: coexiste, gastroentérite, coeffet, etc.

B. The grapheme æ (e.g., cæcum, æquo) is always pronounced /e/; it is transcribed correctly by our rules since it is changed to é in Step 5, page 13. The words which have individual problems (e.g., élæis, ægosome, et cætera) have been added to the lexicon.

Some printers replace æ with é, which -of course- also yields the correct pronunciation (e.g., césium). The only words in which we have seen æ replaced with ae are aequo and vitae (e.g., the 1981 edition of the Petit Larousse Illustré); these two words have been placed in the lexicon.

The "legitimate" occurrences of ae are either treated correctly by our transcription rules (e.g., maestro, maestria) or are included in the lexicon (e.g., Maeterlinck, caennais, de Staël, Saint-Saëns).

APPENDIX XVI: ALGORITHM FOR THE UNSTABLE VOWEL

A. The term unstable vowel designates the sound found in de, le, ceci, monsieur, faisons, etc. It can be dropped or retained depending on complex linguistic, semantic, and social factors. We have selected a style which is less formal than the style used in delivering a speech, but somewhat more formal than the style used in conversation; for example, the following pronunciations -which educated people use only in very informal situations- are not generated by our program:

ʃsɥi for je suis in initial position

lotʒur	for l'autre jour
tabdənɥi	for table de nuit
vunfrerjɛ̃	for vous ne ferez rien

B. For the treatment of the unstable vowel, most books refer to the "loi des trois consonnes" (for example, see Delattre (1949), page 45). This "law" may provide students of French with a satisfactory basis for initial practice, but it is far too simplistic for a computer program. There are several aspects which our program must take into consideration:

1. A search through several dictionaries yielded 79 different initial /CəC/ combinations. Out of these 79 combinations, there are 28 for which we have no evidence that /ə/ is ever dropped: /bəd/, /bəl/, /fəm/, /gən/, /mər/, /pən/, /kən/, /kər/, /səs/, etc. (bedon, belette, femelle, guenon, merisier, penaud, quenouille, querelle, ceci, etc.). There are 32 combinations in which /ə/ is dropped or retained depending on the environment: /dəb/, /dəd/, /dəs/, /fər/, /məz/, /rəm/, /səm/, etc. (debout, dedans, dessous, ferai, mesure, remords, semence, etc.). In the remaining 19 combinations, the fall or retention of /ə/ seems to depend on individual factors; in particular, it seems that the fall of /ə/ depends on the frequency with which words are used in conversation. Examples:

	Retained	Can be dropped
/dəv/	devers	deviens
/fən/	fenouil	fenêtre
/fəz/	faisan	faisant
/mən/	menuet	menu
/pəl/	pelisse	pelouse
/pət/	peton	petit
/rəs/	ressac	ressort
/rət/	retable	retard
/rəv/	revêche	revanche
/sək/	sequin	secret
/ʃəv/	chevesne	cheveu
/vəl/	velu	velours

2. When /ə/ appears in consecutive syllables, the tendency is to select the combination of consonants which is easiest to pronounce:

vous devenez	/vudəvne/	not	/vudvəne/
vous recevez	/vursəve/	not	/vurəsve/
ma chevelure	/maʃəvlyr/	not	/maʃvəlyr/
enseveli	/ɑ̃səvli/	not	/ɑ̃svəli/
il ressemelait	/ilrəsəmlɛ/	not	/ilrəsməlɛ/

3. The syllabic count of the word following an unstable vowel also seems to affect its fall or retention. Examples:

garde-boue	/gardəbu/	(/ə/ kept)
garde-barrière	/gardbarjɛr/	(/ə/ dropped)
porte-clé	/pɔrtəkle/	(/ə/ kept)
porte-monnaie	/pɔrtmɔnɛ/	(/ə/ dropped)

4. In words beginning with re, the tendency to drop /ə/ is greater when re is felt to be a part of the root (as in regretter) than when it is felt to be a prefix (as in regarnir). This appears to be especially true when re is followed by two or three consonants (rechristianiser, restructurer).

5. The unstable vowel, the liaisons, and the distribution of /a-ɑ/ are features of the French language about which educated speakers rarely totally agree. For the unstable vowel, we have attempted to devise a scheme which most educated speakers would find generally acceptable.

C. We transcribe as /ə/ the unstable vowels which, in our opinion, should be retained in the delivery style we have chosen. The unstable vowels which are to be analyzed later are temporarily transcribed as /ɸ/ (whether they should be dropped altogether or changed to /ə/ is decided in Part VII(G), page 76). Our code for the unstable vowels uses the lexicon, the prefix table, the root table, and the transcription rules.

1. To transcribe initial re (or re following a prefix) as /rə/, we use the lexicon (e.g., revis), or the root table (e.g., restatue), or we allow prefix separation (e.g., rebec). The transcription /rɸ/ is obtained by using the lexicon (e.g., repas), or the root table (e.g., regard), or by preventing prefix separation (e.g., reçu since re is followed by only two characters); if the lexicon is used and if the string after re has more than two characters, all the forms in which we want /rɸ/ must be listed (e.g., renard, renards).

The lexicon contains some words which our rules alone could transcribe correctly with /rə/, but we had to place these words in the lexicon since they belong to different grammar categories from those of the unprefixed words (e.g., bouche/rebouche, paies/repaies) or since they resolve ambiguity (e.g., reportions, replacer).

2. The transcription rules handle chevel+V, deven+V, eli+V, eni+V, ensevel+V, eri+V.

3. The transcription rules use /ə/ for the initial strings bed+V, bel+V, ben+V, bes+V, chevr, degr, den+V, faisa (except faisant), faiseu, fenouil, levr, mer+V, pel+V (except pelo, pelu), pen+V, pes+V, quen+V, quer+V, sev, ved+V. In other initial syllables (chev+V, ded+V, dessous, dessus, fer+V, mes+V, pet+V, sem+V, ser+V, etc.), we use /ɸ/ except in these words which have been added to the lexicon:

ceci	cela	celui	chenal	chenaux
chenu	chevêche	chevesne	chevêtre	cheviotte
deça	decauville	dehors	derechef	devers
felouque	femelle	genèse	genevois	genevoise
genèvrier	genèvrière	guenille	guenon	lebel
ledit	menuet	monsieur	peton	selon
sequin	velu			

4. The transcription rules remove the unstable vowel when it precedes a vowel (e.g., seoir).

5. The transformation rules remove the unstable vowel when it follows a vowel (e.g., tuerez).

6. On the list of idioms, we use /ə/ for the first vowel in cases such as de crainte que, de bric et de broc, le leur.

7. All remaining unstable vowels (événement, lentement, vivre, etc.) are transcribed as /ɸ/.

APPENDIX XVII: DESCRIPTION OF LINGUISTIC FEATURES 77 - 112

This appendix contains the lists of words which receive linguistic features 77 through 112.

77.	ce	elle	elles	tu	
78.	la	le	l'	les	
79.	en	leur	lui	moi	m'
	nous	toi	t'	vous	y
80.	de	d'			
81.	le	les			
82.	à				
83.	jamais	pas	plus	rien	
84.	me	m'	nous	se	s'
	te	t'	vous		
85.	y				
86.	en				
87.	leur	lui			
88.	en	la	le	l'	les
	leur	lui	me	m'	nous
	se	s'	te	t'	vous
	y				
89.	à	afin de	afin d'	après	avant de
	avant d'	de	d'	faute de	faute d'
	loin de	loin d'	par	pour	près de
	près d'	quitte à	sans	sauf	
90.	me	m'	se	s'	te
	t'				
91.	lui				

92. en le la l' les
leur nous vous

93. afin de afin d' avant de avant d' loin de
loin d' près de près d'

94. aucun aucune ce certain certaine
certaines certains la le l'
les leur nul nulle plusieurs
quelque son ton tous tout
toute toutes

95. aucun aucune ce certain certaine
certaines certains la le l'
les leur nul nulle plusieurs
quelque

96. aucun aucune aucunes aucuns ça
ce certaines certains c' ç'
elle elles nous nul nulle
personne plusieurs tous tout toutes
tu vous

97. All occurrences of de also represent d'.

à
à cause de,du,des
à côté de,du,des
à défaut de,du,des
à l'abri de,du,des
à la faveur de,du,des
à la merci de,du,des
à l'égard de,du,des
après
au-dedans de,du,des
au-dehors de,du,des
au-delà de,du,des
au-dessous de,du,des
au-dessus de,du,des
au-devant de,du,des
au lieu de,du,des
au milieu de,du,des
auprès de,du,des
autour de,du,des
au travers de,du,des
aux dépens de,du,des
avant
avec
chez
comme
contre
dans
de
de chez
d'entre
depuis
derrière
dès
devant
en
en bas de,du,des
en dépit de,du,des
en face de,du,des
en faveur de,du,des
en haut de,du,des
entre
envers
face à,au,aux
faute de
grâce à,au,aux
hors
hors de,du,des
jusqu'à,au,aux,en
loin de,du,des
lors de,du,des
malgré
outre
par
par-dessous
par-dessus
par-devant
parmi
par rapport à,au,aux
pendant
pour
près de,du,des
proche de,du,des
quant à,au,aux
sans
selon
sous
sur
vers
vis-à-vis de,du,des

98. elle elles eux lui moi
nous soi toi vous

99. l' m' t'

100.	leur	lui	me	m'	nous
	se	s'	te	t'	vous
	y				
101.	les	lui	y		
102.	à	après	avant	avec	chez
	comme	contre	dans	de	d'
	derrière	dès	devant	envers	malgré
	outre	par	pendant	pour	sans
	sauf	selon	sous	sur	vers
103.	au	aucun	aucune	ce	cet
	cette	chaque	du	mon	notre
	son	ton	tout	toute	un
	une	votre			
104.	aucunes	aucuns	celles	celles-ci	
	celles-là	certaines	certains	ceux	
	ceux-ci	ceux-là	elles	ils	
	les leurs	les miennes	les miens	les nôtres	
	les siennes	les siens	les vôtres	lesquelles	
	lesquels	leurs	plusieurs	quelques-unes	
	quelques-uns	qui	tous	toutes	
105.	le	la	l'	leur	en
106.	aucun	aucune	ça	ceci	cela
	celle	celle-ci	celle-là	celui	celui-ci
	celui-là	certain	certaine	elle	il
	la leur	la mienne	la nôtre	la sienne	la vôtre
	laquelle	le leur	le mien	le nôtre	le sien
	le tien	le vôtre	lequel	nul	nulle
	on	personne	quelqu'un	quelqu'une	rien
	toute	un	une		
107.	et	ni			
108.	certain	certaine	la	le	l'
	leur	nul	nulle	quel	quelle
109.	auquel	auxquelles	auxquels	dont	où
	quoi				
110.	que	qu'			
111.	auquel	auxquelles	auxquels	desquelles	desquels
	dont	duquel	où	que	qu'
	qui				

112. The items listed in 112 are used to analyze words which can be determiners and pronouns. The string de cannot be replaced with d'.

à côté de	à l'égard de	à travers	auprès de
autour de	au travers de	au-dedans de	au-dehors de
au-delà de	au-dessous de	au-dessus de	au-devant de
aux dépens de	avant	avec	chez
contre	dans	de chez	d'entre
depuis	derrière	devant	dès
en	en bas de	en face de	en haut de
envers	face à	grâce à	hors
hors de	lors de	malgré	parmi
par rapport à	par-dessous	par-dessus	par-devant
pendant	proche de	selon	sous
sur	vers	vis-à-vis de	voici
voilà			

APPENDIX XVIII: FIGURES

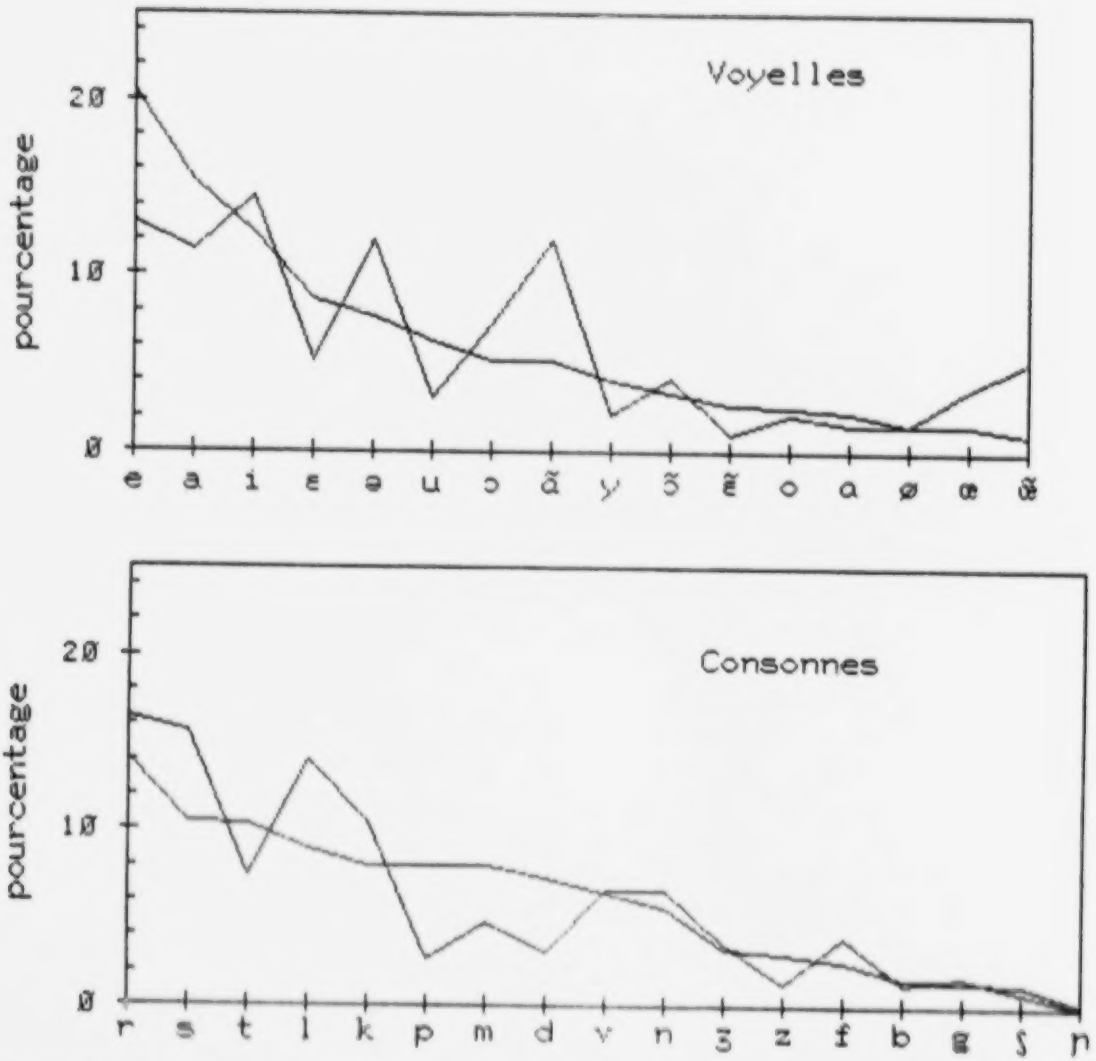

Figure 1: The jagged lines represent the distribution of phonemes in Baudelaire's Harmonie du soir. The gradual lines represent the distribution found in conversation.

FLOWCHART FOR THE TRANSCRIPTION OF INDIVIDUAL WORDS

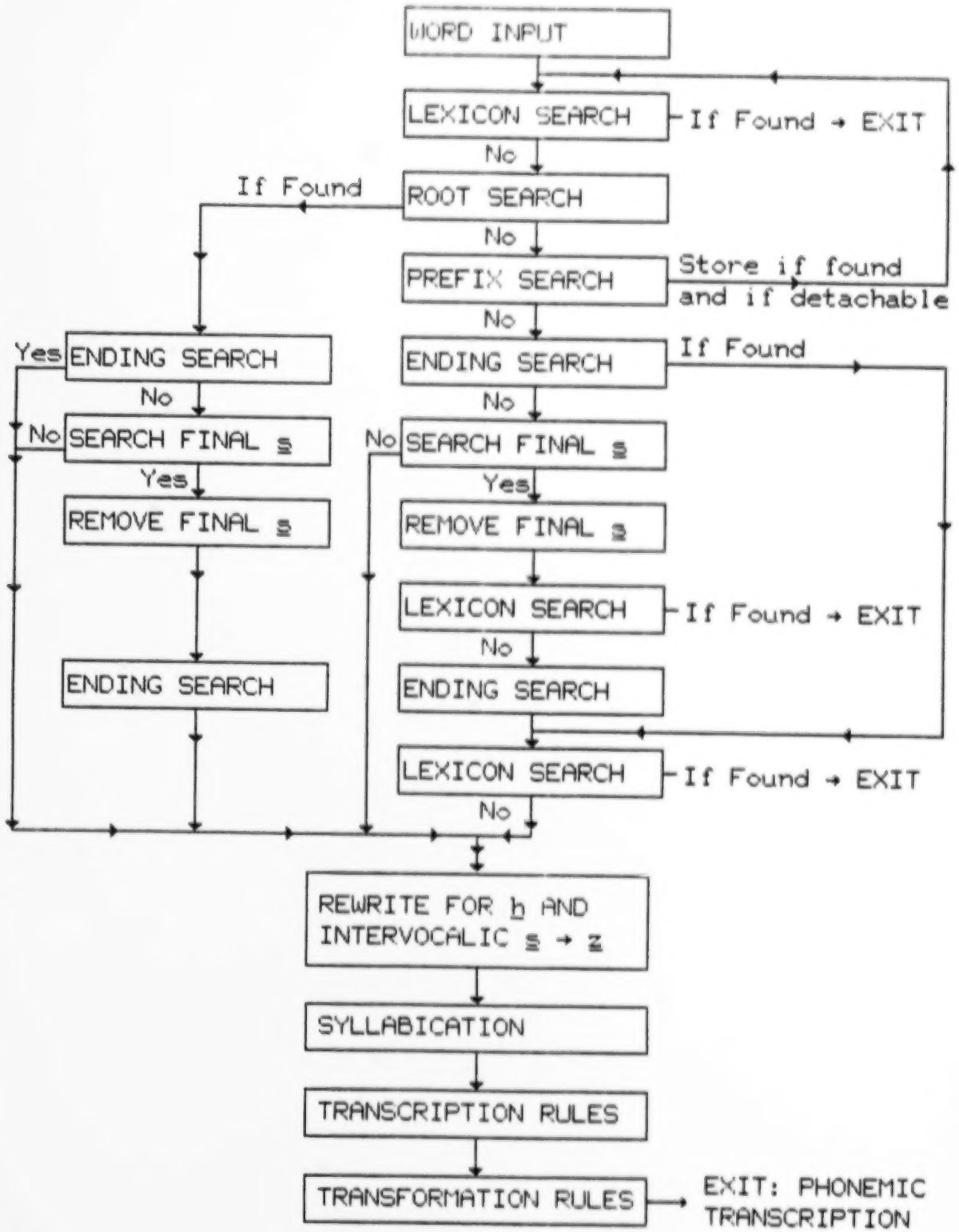

Figure 2

Sequence of steps followed by word "macrosuperdésulfatations" as it proceeds through the phonemic transcriber.

```
Search lexicon:      macrosuperdésulfatations   NOT FOUND
Search root list:    macrosuperdésulfatations   NOT FOUND
Search prefix list:  macrosuperdésulfatations   FOUND
     Transcribe /macro/, remove and store
Search lexicon:      superdésulfatations        NOT FOUND
Search root list:    superdésulfatations        NOT FOUND
Search prefix list:  superdésulfatations        FOUND
     Transcribe /syper/, remove and store
Search lexicon:      désulfatations             NOT FOUND
Search root list:    désulfatations             FOUND
     Rewrite: déçulfatations
Search ending list:  déçulfatations             FOUND
     Rewrite: déçulfatassion

Transcription rule:      "d"          →   /d/
Transcription rule:      "é"          →   /e/
Transcription rule:      "ç"          →   /s/
Transcription rule:      "u"          →   /y/
Transcription rule:      "l"          →   /l/
Transcription rule:      "f"          →   /f/
Transcription rule:      "a"          →   /a/
Transcription rule:      "t"          →   /t/
Transcription rule:      "a"          →   /a/
Transcription rule:      "ss"         →   /s/
Transcription rule:      "i"          →   /i/
Transcription rule:      "final on"   →   /3/
Re-attach prefixes /makro/ and /syper/
Transformation rule:     /i/ in one C+/i/+V     →  /j/
Transcription:           /makrosyperdesylfatasj3/
```

Figure 3

FOOTNOTES

1. This includes short stories, novels, essays, plays, news magazines, newspapers (excluding classified ads), and textbooks about subjects such as history, geography, and literature; our program is designed to handle the few mathematical and scientific strings which are likely to occur in such narrations or dialogues. However, our program is not designed to transcribe texts dealing specifically with physics, chemistry, mathematics, economics, contemporary logic, or any topic making use of specialized symbols or characters: transcribing such texts presents an entirely different set of problems and requires an entirely different approach.

Our program will read poetry as if it were prose; this is hardly objectionable for most modern poets (e.g., Prévert), but it is not suitable for classical tragedies (e.g., Racine's Phèdre), romantic poetry (e.g., Lamartine's Le Lac), etc. A proper rendition of such works requires a different treatment of liaisons, unstable vowels, and semivowels. We attempted to add to our program an option which would have allowed the user to choose "prose" or "poetry", but the necessary modifications ran so deep into the program that its size would have nearly doubled. We believe it is preferable to write a separate program for poetry reading.

Our program assumes that the text uses the spelling adopted by the French Academy in the eighth edition (1935) of its dictionary; it also accepts the changes made in 1976 (see page 33). It makes only very few errors with texts using the spelling of the seventh edition (1878) or of the sixth edition (1835), but more and more errors occur the further back we go; for example, in a randomly chosen passage from Rabelais' Pantagruel printed with the spelling used at that time (first fourteen paragraphs of Chapter XXV), our program is unable to handle 48 of the 307 words (e.g., estoit, chevaulx, cependent, congnoistrez, aultres).

Involuntary spelling errors (i.e., those made by uneducated people), depending on type and frequency, affect the accuracy of this program in different ways; voluntary spelling errors (for example, those found in Queneau's Zazie dans le métro) yield a correct phonemic transcription (e.g., lagoçamilébou, singermindépré) more often than not (e.g., imfõ, cexé), but the linguistic analysis is wrong.

2. We mean fourteen syllables before the unstable vowels which can be dropped are removed. In the final transcription, the average length of a breath-group is twelve syllables.

3. The "registre" chosen for this program is the one which is normally used by educated speakers in informal conversation: few optional liaisons are made and most of the unstable vowels which can be dropped are dropped. Thus, the sentence:

Je ne lis pas de livres anglais.

is transcribed /ʒ ɔnlipɑdlivrɑ̃glɛ/, and not /ʒənəlipɑdəlivrəzɑ̃glɛ/.

4. Phonemic accuracy can be easily measured since a standard can be established and individual phonemes can be marked right or wrong. Prosody combines frequency, intensity, and length; the rules which produce the prosodic curves are necessarily complex since they require data about the grammar categories, the nature of the phonemes and their environment, the punctuation marks, the changes in typeface, etc.

At the present time, it does not seem possible to prepare a computer program which would analyze the emotional content of a text and produce prosodic curves expressing contempt, jealousy, hatred, anger, love, etc.

5. Even the best optical scanners require that the characters be clearly formed and sharply printed. A letter o broken on the right side is read as a c and, for example, oiseler is picked up as ciseler. A human reader would know, from the context, that the word has to be oiseler, but there does not seem to be any possibility that a computer program will ever be able to perform the enormous number of contextual analyses that would be necessary to prevent such misreadings; since a poorly-printed character can occur anywhere, the program would have to ascertain —as human readers do unconsciously— that every word fits the context.

6. Professor Bruce Sherwood (1981) writes:

"Of critical significance is the capability to generate new messages by the algorithmic creation of character strings. This makes text-to-speech a much more general output scheme than either analog or digital recording. In the long run, this aspect is likely to make text-to-speech the preferred approach in computer-based education if the quality ever reaches adequately high levels at adequately low costs."

7. Words which have an accent on their initial letter (e.g., âge, émouvoir, être, ôter) present a problem when the initial letter is capitalized and the accent is not used (e.g., Emouvoir les foules n'est pas chose facile, la Caisse d'Epargne). We address that problem as follows:

a. We have a transcription rule which states that initial e followed by a syllabic cut is to be transcribed /e/. The words êtes, être, êtres are in the lexicon with and without an accent since they use /ɛ/.

b. We have a transcription rule which states that initial ota, ote, and oti are to be transcribed with /o/ and we have added these words to our lexicon: otage, otalgie, otarie, otique, otite (with /ɔ/), and otons (with /o/).

c. All the words in our lexicon which begin with é have also been listed without the accent. A search through the Petit Robert (1979) indicates that the only other words which need to be listed are age for âge, agé for âgé, ame for âme, atre for âtre. There is no need to list the words in which the absence of the accent does not change the pronunciation (e.g., île).

d. The prefix électro is also listed as electro in the prefix table so that words prefixed with Electro will receive their correct features

(e.g., Electroménager) and their correct pronunciation (Electropuncture).

Headlines written entirely in capital letters without accents would present a serious problem if they were frequent. In the following examples:

UN COUTURIER NOMME BORG

LE POSTE PREFERE DES FEMMES

our program could not determine whether the writer means "nomme" or "nommé" and "préféré" or "préfère". Fortunately, in such headlines, accents are generally used when ambiguity could exist, whether the pronunciation is different or not (DU / DÛ, LA / LÀ).

8. In Proust's novels, for example. There are also some writers who have written books without any punctuation at all (e.g., Comment c'est by Samuel Beckett, 1961, Les éditions de minuit).

9. Treating ellipsis marks as true punctuation is acceptable in cases such as:

Il m'a dit M..... bien calmement.

but it creates unnatural breath-groups in cases such as:

Le 17 juin 1984, D... B... avait au cours d'une dispute poignardé sa femme.

10. The decision to stop using periods to separate units from thousands, thousands from millions, etc. was made in 1935. The old system has not totally disappeared (e.g., 12.964.287), but most people now use a space instead of a period (e.g., 12 964 287); in dates (e.g., 1986), nothing is used. Our program accepts all those forms. A period which precedes only one or two figures is read as point (e.g., 13.7 becomes treize point sept), but in a sequence "number lower than 31"."number lower than 12"."number greater than 1000 and lower than 2100", the second number is replaced with the corresponding graphemic form of the month (e.g., janvier, février).

11. Dubois (1977) lists over 25,000 abbreviations in his dictionary.

12. We have not included the abbreviations found in classified ads; example:

MAISON F6 parf. ét. gar. 2 voit.
chff. cent. gaz, 1.000 m^2 terr.,
3.000 culture. 200.000 F cpt. Loy.
150.000 F. (43) 89.88.35 ap. 20 h.

JF. Bl., yx bl., née Tours, hab. RP,
ch. JH, 1,65 env., chx ch., Corse,
connu dim. 24 juin, place Bastille,
angl. Beaumarchais.
(Le Nouvel Observateur, July 6, 1984)

13. However, this is not always true. Example:

> Réponse courte: un seul mot, mais en lettres épaisses qui emplissent la largeur de la carte de visite: MERDE.
> (Françoise Dorin, Les lits à une place. Paris: Flammarion, 1980, page 13)

To prevent this word and a few other words we have found in our corpus from being treated like abbreviations and read letter by letter, we have added them to Group 6 of the lexicon.

14. We use the term "explicit" to designate the first eighteen linguistic features (page 188) because they are applied to words which can belong to only one grammar category (e.g., chez, nouvel). We use the term "ambiguous" for features 19 through 51 since they are applied to words which belong to several grammar categories (e.g., court, plus).

15. For example, Le Canard Enchaîné (July 25, 1984) has this title:

> Mitterrand tire les fils ... spirituels

16. There have been considerable changes in the use of liaisons during the past century. Many liaisons which Rousselot (1902) listed as compulsory are no longer made. In particular, the liaison with /k/ survives in only a few examples which we have placed in our list of idioms (e.g., franc-alleu, respect humain, sang et eau); with the adjective long, practically everyone today pronounces a /g/ of liaison: /œlɜ̃givɛr/.

17. But, even with the list of idioms, there are still a few ambiguous constructions which our program cannot handle with total accuracy. For example, we do not always analyze tous correctly; our program transcribes pour tous les mêmes travaux as /purtulemɛmtravo/ although, in some contexts, it should be /purtuslemɛmtravo/.

18. Highly technical articles may pose problems; for example, a newspaper article discussing the value of stocks on the New York Stock Exchange will contain many proper names which our program will transcribe in an unacceptable way (that is, a pronunciation which educated French persons would not use).

19. In some cases, é is used (e.g., cénure); this spelling —of course— poses no problem.

B I B L I O G R A P H Y

ALLEN J. (1976). Synthesis of speech from unrestricted text. Proceedings of the Institute of Electric and Electronic Engineers, 64, 433-442.

ASTA V., LIÉNARD J.S. (1979). L'ICOPHONE logiciel: un synthétiseur par forme d'ondes. 10e journées d'étude sur la parole. Groupement des acousticiens de langue française, Grenoble.

CATACH N. (1971). Orthographe et Lexicographie (Tome 1). Paris: Didier.

CATACH N. (1981). Orthographe et Lexicographie (Tome 2). Paris: Nathan.

CELLARD J., REY A. (1980). Dictionnaire du français non-conventionnel. Paris: Hachette.

DELATTRE P. (1947). Principes de phonétique française à l'usage des étudiants anglo-américains. Middlebury, Middlebury College (Vermont).

DELATTRE P. (1949). Advanced Training in French Pronunciation. Middlebury, Middlebury College (Vermont).

DELATTRE P. (1957). La question des deux "a" en français. French Review, XXXI, 141-148.

DELATTRE P. (1966). Les dix intonations de base du français. French Review, XL, 1-14.

DESFEUILLES P.(1961). Dictionnaire de rimes. Paris: Garnier Frères.

DIVAY M. & GUYOMARD M. (1977). Conception et réalisation sur ordinateur d'un système de transcription graphémo-phonétique du français. Thèse de Troisième Cycle, Université de Rennes.

DIVAY M. & GUYOMARD M. (1977). Grapheme to phoneme transcription for French. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Hartford, 575-578.

DUBOIS M. (1977). French and International Acronyms and Initialisms Dictionary. New York: Marlin.

ELOVITZ H.S., JOHNSON R., McHUGH A. & SHORE J.E. (1976). Letter-to-sound rules for automatic translation of English Text to Phonetics. IEEE Transaction on Acoustics, Speech, and Signal Processing, ASSP-24, 446-459.

ÉMERARD F., LARREUR D. (1976). Synthèse par diphones et traitement automatique de la prosodie. Compte rendu des 7e journées d'études sur la parole, Nancy, 179-192.

FERVERS H., MICLET L., LE ROUX J. Programme de transcription orthographique phonémique en langue française. E.N.S.T. - D - 76 003, E.N.S.T, Paris.

FOUCHÉ P. (1956). Traité de prononciation française. Paris: Klincksieck.

FRÉMY D. & M. (1981). QUID. Paris, Robert Laffont.

GAUVAIN J-L., GANGOLF J-J. (1983). Terminal integrates speech recognition and text-to-speech synthesis. Speech Technology, Sept./Oct. 1983, 25-38.

GRESSER J.Y. (1973). Passage de l'orthographe à une transcription phonétique. Compte rendu des recherches menées au département CSI, no. 1, February 1973.

GREVISSE M. (1980). Le Bon Usage. Paris-Gembloux: Duculot.

GUYOMARD M. (1974). Transcription du français en phonétique. Compte rendu des recherches menées au département CSI, no. 2, February 1974.

HUG M. (1979). La distribution des phonèmes en français. Genève: Slatkine.

JUILLAND A. (1965). Dictionnaire inverse de la langue française. The Hague: Mouton and Co.

LAROUSSE (1981). Petit Larousse Illustré. Paris: Librairie Larousse.

LE CORNEC A. La transcription orthographique-phonétique du français. Note CEI/CSI/22, CNET, Lannion, France.

LEROND A. (1980). Dictionnaire de la prononciation. Paris: Larousse.

LIÉNARD S.C., TEIL D., CHOPPY C., RENARD G., & SAPALY J. (1977). Diphone synthesis of French: vocal response unit and automatic prosody from the text. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Hartford, 560-563.

LUCCI, V. (1983). Etude phonétique du français contemporain à travers la variation situationelle: débit, rythme, accent, intonation, (schwa) muet, liaisons, phonèmes. Grenoble: Université des langues et lettres.

LYNE A. (1973). L'élaboration des listes de fréquence. A la recherche d'une solution aux problèmes d'affectation des mots-occurrences dans les classes de mots. Cahiers de lexicologie, 23 (1973-2), 83-108.

MAGGS P., TRESCASES P. (1980). De l'écrit à l'oral: un programme sur micro-ordinateur pour machine à parler à l'usage des aveugles francophones. Le Français Moderne, 48, no. 3, 225-245.

MARTIN Ph. (1976). Synthèse par règles de l'intonation de la phrase. Compte rendu des 7e journées d'études sur la parole, Nancy, 205-213.

MARTINET A., WALTER H. (1973). Dictionnaire de la prononciation française dans son usage réel. Paris: France-Expansion.

MARTINON Ph. (1962). Dictionnaire des rimes françaises. Paris: Larousse.

McILROY M.D. (1974). Synthetic English speech by rule. Computing Science Technical Report 14, Bell Laboratories.

OAKEY S., CAWTHORN R.C. (1981). A rule learning system and its application to machine speech. Proceedings of Acoustics, Spring 1981: 295-298.

OAKEY S., CAWTHORN R.C. (1981). Inductive learning of pronunciation rules by hypothesis testing and correction. IJCAI-81 Proceedings: 109-114.

PROUTS B. (1979). Traduction phonétique de textes écrits en français. 10e journées d'étude sur la parole. Groupement des acousticiens de langue française, Grenoble.

QUENEAU R. (1959). Zazie dans le métro. Paris: Gallimard.

QUID (1981), see Frémy

RABELAIS F. (1922). Œuvres de François Rabelais (Tome IV). Paris: Champion.

ROBERT P. (1979). Dictionnaire alphabétique et analogique de la langue française. Paris: Société du Nouveau Littré.

RODET X., POIROT L. Transcription phonétique automatique du français. SES/INTERNE/SERF/76-068.

ROUSSELOT (Abbé), LACLOTTE F. (1902). Précis de Prononciation Française. Paris: Welter.

SHERWOOD B. A. (1978). Fast text-to-speech algorithms for Esperanto, Spanish, Italian, Russian and English. International Journal of Man-Machine Studies, 10, 669-692.

SHERWOOD B. A. (1981). New technology provides computer voices for education. Speech Technology, I, 1 (Fall 1981): 24-29.

SHERWOOD B. A. and J. N. SHERWOOD (1982). Computer voices and ears furnish novel teaching options. Speech Technology, September/October 1982, 46-51.

SILVA G. An automatic orthographic-to-phonetic conversion system for French. Institute of Library Research, University of California, Los Angeles.

TEJA E. R. (1981). Teaching your computer to talk. Blue Ridge Summit, Pa: Tab Books Inc.

UMEDA N. (1976). Linguistic rules for text-to-speech synthesis. Proceedings of the IEEE, 64, 443-451.

VAISSIÈRE J. (1976). Une procédure de segmentation automatique de la parole en mots prosodiques en français. Compte rendu des 7e journées d'études sur la parole, Nancy, 105-114.

WARNANT L. (1962). Dictionnaire de la prononciation française. Gembloux: Éditions Duculot.

WITTEN I.H. (1976). Generating natural speech from text. Proceedings AISB Summer Conference, Edinburgh, July.